Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data

Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data

Nils Eling[1,2], Arianne C. Richard[2,3], Sylvia Richardson[4], John C. Marioni[1,2,*] and Catalina A. Vallejos[5,6,7]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[2] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, CB2 0RE, UK

[3] Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0XY, UK

[4] MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK

[5] The Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, UK

[6] Department of Statistical Science, University College London, 1-19 Torrington Place, London WC1E 7HB, UK

[7] MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XY, UK

* Corresponding author and lead contact: marioni@ebi.ac.uk

Summary

Cell-to-cell transcriptional variability in otherwise homogeneous cell populations plays a crucial role in tissue function and development. Single-cell RNA sequencing can characterise this variability in a transcriptome-wide manner. However, technical variation and the confounding between variability and mean expression estimates hinders meaningful comparison of expression variability between cell populations. To address this problem, we introduce an analysis approach that extends the BASiCS statistical framework to derive a residual measure of variability that is not confounded by mean expression. This includes a robust procedure for quantifying technical noise in experiments where technical spike-in molecules are not available. We illustrate how our method provides biological insight into the dynamics of cell-to-cell expression variability, highlighting a synchronisation of biosynthetic machinery components in immune cells upon activation. In contrast to the uniform up-regulation of the biosynthetic machinery, $CD4^+$ T cells show heterogeneous up-regulation of immune-related and lineage-defining genes during activation and differentiation.

Introduction

Heterogeneity in gene expression within a population of single cells can arise from a variety of factors. Structural differences in gene expression within a cell population can reflect the presence of sub-populations of functionally different cell types (Zeisel et al., 2015; Paul et al., 2015). Alternatively, in a seemingly homogeneous population of cells, so-called unstructured expression heterogeneity can be linked to intrinsic or extrinsic noise (Elowitz et al., 2002). Changes in physiological cell states (such as cell cycle, metabolism, abundance of transcriptional/translational machinery and growth rate) represent extrinsic noise, which has been found to influence expression variability within cell populations (Keren et al., 2015; Buettner et al., 2015; Zeng et al., 2017). Intrinsic noise can be linked to epigenetic diversity (Smallwood et al., 2014), chromatin rearrangements (Buenrostro et al., 2015), as well as the genomic content of single genes, such as the presence of TATA-box motifs and the abundance of nucleosomes around the transcriptional start site (Hornung et al., 2012).

Single-cell RNA sequencing (scRNAseq) generates transcriptional profiles of single cells, allowing the study of cell-to-cell heterogeneity on a transcriptome-wide (Grün et al., 2014) and single gene level (Goolam et al., 2016). Consequently, this technique can be used to study unstructured cell-to-cell variation in gene expression within and between homogeneous cell populations (i.e. where no distinct cell sub-types are present). Increasing evidence suggests that this heterogeneity plays an important role in normal development (Chang et al., 2008) and that control of expression noise is important for tissue function (Bahar Halpern et al., 2015). For instance, molecular noise was shown to increase before cells commit to lineages during differentiation (Mojtahedi et al., 2016), while the opposite is observed once an irreversible cell state is reached (Richard et al., 2016). A similar pattern occurs during gastrulation, where expression noise is high in the uncommitted inner cell mass compared to the committed epiblast and where an increase in heterogeneity is observed when cells exit the pluripotent state and form the uncommitted epiblast (Mohammed et al., 2017).

Motivated by scRNAseq, recent studies have extended traditional differential expression analyses to explore more general patterns that characterise differences between cell populations (e.g. Korthauer et al., 2016). In particular, BASiCS (Vallejos et al., 2015, 2016) introduced a probabilistic tool to assess differences in cell-to-cell heterogeneity between two or more cell populations. This feature has led to, for example, insights into the context of immune activation and ageing (Martinez-Jimenez et al., 2017). To meaningfully assess changes in biological variability across the entire transcriptome, two main confounding effects must be taken into account: differences due to artefactual technical noise and dif-

ferential variability between populations that is driven by changes in mean expression. The latter arises because biological noise is negatively correlated with protein abundance (Bar-Even et al., 2006; Newman et al., 2006; Taniguchi et al., 2010) or mean RNA expression (Brennecke et al., 2013; Antolović et al., 2017). To address these two confounding effects, BASiCS separates biological noise from technical variability by borrowing information from synthetic RNA spike-in molecules. Additionally to acknowledge the variance-mean relationship, it restricts differential variability testing to those genes with equal mean expression across populations.

This article extends the statistical model implemented in BASiCS by implementing a more general approach to account for the aforementioned confounding effects. Firstly, we derive a residual measure of cell-to-cell transcriptional variability that is not confounded by mean expression. This is used to define a probabilistic rule to robustly highlight changes in variability, even for differentially expressed genes. Unlike previous related methods (e.g. Kolodziejczyk et al., 2015), our approach directly performs gene-specific statistical testing between two conditions using a readily available measure of uncertainty. Secondly, by exploiting concepts from measurement error models, our method is extended to address experimental designs where spike-in sequences are not available. This is particularly critical due to the increasing popularity of droplet-based technologies.

Using our approach, we identify a synchronisation of biosynthetic machinery components in $CD4^+$ T cells upon early immune activation as well as an increased variability in the expression of genes related to $CD4^+$ T cell immunological function. Furthermore, we detect evidence of early cell fate commitment of $CD4^+$ T cells during malaria infection characterized by a decrease in *Tbx21* expression heterogeneity and a rapid collapse of global transcriptional variability after infection. These results highlight biological insights into T cell activation and differentiation that are only revealed by jointly studying changes in mean expression and variability.

Results

Addressing the mean confounding effect for differential variability testing

Unlike bulk RNA sequencing, scRNAseq provides information about cell-to-cell expression heterogeneity within a population of cells. Past works have used a variety of measures to quantify this heterogeneity. Among others, this includes the coefficient of variation (CV, Brennecke et al., 2013) and entropy measures (Richard et al., 2016). As in Vallejos et al. (2015, 2016), we focus on biological *over-dispersion* as a proxy for transcriptional heterogeneity. This is defined by the excess of variability that is observed with respect to what would be predicted by Poisson sampling noise, after accounting for technical variation.

The aforementioned measures of variability can be used to identify genes whose transcriptional heterogeneity differs between groups of cells (defined by experimental conditions or cell types). However, the strong relationship that is typically observed between variability and mean estimates (e.g. Brennecke et al., 2013) can hinder the interpretation of these results.

A simple solution to avoid this confounding is to restrict the assessment of differential variability to those genes with equal mean expression across populations (see **Figure 1A**). However, this is sub-optimal, particularly when a large number of genes are differentially expressed between the populations. For example, reactive genes that change in mean expression upon changing conditions (e.g. transcription factors) are excluded from differential variability testing. An alternative approach is to directly adjust variability measures to remove this confounding. For example, Kolodziejczyk et al. (2015) computed the empirical distance between the squared CV to a rolling median along expression levels — referred to as the DM method.

In line with this idea, our method extends the statistical model implemented in BASiCS (Vallejos et al., 2015, 2016). We define a measure of *residual over-dispersion* — which is not correlated with mean expression — to meaningfully assess changes in transcriptional heterogeneity when genes exhibit shifts in mean expression (see **Figure 1B**). More concretely, we infer a regression trend between over-dispersion (δ_i) and gene-specific mean parameters (μ_i), by introducing a joint informative prior to capture the dependence between these parameters (see **STAR Methods**). A latent gene-specific *residual over-dispersion* parameter ϵ_i describes departures from this trend (see **Figure 1C**). Positive values of ϵ_i indicate that a gene exhibits more variation than expected relative to genes with similar

expression levels. Similarly, negative values of ϵ_i suggest less variation than expected and, as shown in **Figure 1D**, these residual over-dispersion parameters are not confounded by mean expression.

Our hierarchical Bayes approach infers full posterior distributions for the gene-specific latent residual over-dispersion parameters ϵ_i. As a result, we can directly use a probabilistic approach to identify genes with large absolute differences in residual over-dispersion between two groups of cells (see **Figure 1E** and **STAR Methods**). The performance of this differential variability test was validated using simulated data (see **Figure S1** and **STAR Methods**). In contrast, mean-corrected point estimates for residual noise parameters (such as those obtained by the DM method) cannot be directly used to perform gene-specific statistical testing between two conditions as no measure of the uncertainty in the estimate is readily available.

The informative prior stabilizes parameter estimation

Our joint prior formulation has introduced a non-linear regression to capture the overall trend between gene-specific over-dispersion parameters δ_i and mean expression parameters μ_i (see **STAR Methods**). Thus, we also refer to the extended model induced by this prior as the *regression* BASiCS model. Accordingly, the model induced by the original independent prior specification (Vallejos et al., 2016) is referred to as the *non-regression* BASiCS model.

To study the performance of the regression BASiCS model, we applied it to a variety of scRNAseq datasets. Each dataset is unique in its composition, covering a range of different cell types and experimental protocols (see **STAR Methods** and **Table S1**). Qualitatively, we observe that the inferred regression trend varies substantially across different datasets (**Figure 2** and **Figure S2**), justifying the choice of a flexible semi-parametric approach (see **STAR Methods**). Moreover, as expected, we observe that residual over-dispersion parameters ϵ_i are not confounded by mean expression, nor by the percentage of zero counts per gene.

The regression BASiCS model introduces a joint prior specification for $(\mu_i, \delta_i)'$, shrinking the posterior estimates for μ_i and δ_i towards the regression trend (this is in line with the shrinkage observed in Love et al., 2014). The strength of this shrinkage is dataset-specific, being more prominent in sparser datasets with a higher frequency of zero counts (see **Figure 2A**) and for lowly-expressed genes where measurement error is greatest.

Subsequently, we asked whether or not the shrinkage introduced by the regression BASiCS model improves posterior inference. To assess this, we compared estimates for gene-specific parameters across: (i) different sample sizes and (ii) different gene expression levels. More concretely, we used a large dataset containing 939 CA1 pyramidal neurons (Zeisel et al., 2015) to artificially generate smaller datasets by randomly sub-sampling 50-500 cells. For each sample size, parameter estimates were then obtained using both the regression and non-regression BASiCS models. The distribution of these estimates is summarised in **Figure 3.**

Firstly, we observe that both the regression and non-regression BASiCS models led to comparable and largely stable mean expression estimates across different sample sizes and expression levels (see **Figure 3A**). Secondly, in line with the results in **Figure 2**, the main differences between the methods arise when estimating the over-dispersion parameters δ_i (see **Figure 3B** and **Figure S3A-C**). In particular, we observe that the non-regression BASiCS model appears to underestimate δ_i for lowly expressed genes when the sample size is small (with respect to the parameter estimates obtained based on the full dataset of 939 cells). In contrast, the shrinkage introduced by our regression BASiCS model aids parameter estimation, leading to robust estimates even for the smallest sample size. This is particularly important for rare cell populations where large sample sizes are difficult to obtain. A similar effect is observed for genes with medium and high expression levels, where the non-regression BASiCS model appears to overestimate δ_i. We also observe that estimates of residual over-dispersion parameters ϵ_i are stable across sample sizes and expression levels. These findings are replicated across multiple sub-sampling experiments (see **Figure S3D-F**).

As an external validation, we compared our posterior estimates of gene-specific model parameters obtained from scRNAseq data to empirical estimates from matched smFISH data of mouse embryonic stem cells grown in 2i and serum media (see **STAR Methods** and Grün et al., 2014). Firstly, posterior estimates of mean-expression parameters μ_i exhibit high correlation to smFISH mean transcript counts (see **Figure S3G**). Secondly, we also observe a strong correlation between posterior estimates for over-dispersion parameters δ_i and the empirical CV^2 values obtained from smFISH data (see **Figure S3H**). Finally, a similar behaviour is observed when comparing posterior estimates of residual over-dispersion parameters ϵ_i to a residual CV^2 (see **Figure S3I** and **STAR Methods**).

Inferring technical variability without spike-in genes

Another critical aspect to take into account when inferring transcriptional variability based on scRNAseq datasets is technical variation (Brennecke et al., 2013). BASiCS achieves this through a vertical data integration approach, exploiting a set of synthetic RNA spike-in molecules (e.g. the set of 92 ERCC molecules developed by Jiang et al., 2011) as a *gold standard* to aid normalisation and to quantify technical artefacts (see **Figure 4A**). However, while the addition of spike-in genes prior to sequencing is theoretically appealing (Lun et al., 2017), several practical limitations can preclude their utility in practice (Vallejos et al., 2017). Furthermore, the use of spike-in genes is not compatible with (increasingly popular) droplet-based technologies which have massively increased the throughput of scRNAseq over the last few years (Svensson et al., 2018).

Consequently, to ensure the broad applicability of our method, we extend BASiCS (both the regression and non regression models) to handle datasets without spike-in genes. For this purpose, we exploit principles of measurement error models where — in the absence of gold standard features — technical variation is quantified through *replication* (Carroll, 1998). As described in **Figure 4B**, this horizontal data integration approach is based on experimental designs where cells from a population are randomly allocated to multiple independent experimental replicates (here referred to as *batches*). In such an experimental design, the no-spikes implementation of BASiCS assumes that biological effects are shared across batches and that technical variation will be reflected by spurious differences. As shown in **Figure 4C-D**, posterior inference under the no-spikes BASiCS model closely matches the original implementation for datasets where spike-ins and batches are available. Technical details about the no-spikes implementation of BASiCS are discussed in **STAR Methods** and **Figure S4**.

Expression variability dynamics during immune activation and differentiation

Here, we illustrate how our method assesses changes in expression variability using $CD4^+$ T cells as model system. For all datasets, pre-processing steps are described in **STAR Methods**.

Testing variability changes in immune response gene expression

To identify gene expression changes during early T cell activation, we compared $CD4^+$ T cells before (naive) and after (active) 3 hours of stimulation (Martinez-Jimenez et al., 2017). When using the non-regression BASiCS model, our differential over-dispersion test

avoided the confounding with mean expression by solely focusing on genes with no changes in mean expression. This represents only a small fraction out of the full set of expressed genes. In contrast, testing changes in variability through residual over-dispersion allows testing across all genes, including the large set of genes that are up-regulated upon immune activation (see **Figure S5A-B** and **STAR Methods**). The latter include immune-response genes and critical drivers for $CD4^+$ T cell functionality.

Our model classifies genes into four categories based on their expression dynamics: down-regulated upon activation with (i) lower and (ii) higher variability, and up-regulated with (iii) lower and (iv) higher variability (see **Figure 5A**, **STAR Methods** and Table S2).

Genes with up-regulated expression upon activation and decreased expression variability encode components of the splicing machinery (e.g. *Sf3a3*, *Plrg1*), RNA polymerase subunits (e.g. *Polr2l*, *Polr1d*) as well as translation machinery components (e.g. *Ncl*, *Naf1*) (see **Figure 5B**). These biosynthetic processes help naive T cells to rapidly enter a program of proliferation and effector molecule synthesis (Tan et al., 2017; Araki et al., 2017). Therefore, rapid, uniform up-regulation of these transcripts would assist such processes. This observation also confirms previous findings that the translational machinery is tightly regulated during early immune activation (Martinez-Jimenez et al., 2017).

In contrast, genes with up-regulated expression and increased expression variability (see **Figure 5C**) include the death-inducing and inhibitory transmembrane ligands Fas ligand (*Fasl*) and PD-L1 (*Cd274*), the regulatory transcription factor Smad3 (*Smad3*), and the TCR-induced transcription factor, Oct2 (*Pou2f2*). Additionally, we detect a heterogeneous up-regulation in the mRNA expression of the autocrine/paracrine growth factor IL-2 (*Il2*) upon immune activation. This is in line with previous reports of binary IL-2 expression within a population of activated T cells, which has been suggested to be necessary for a scalable antigen response (Fuhrmann et al., 2016). Heterogeneity in expression of these genes suggests that, despite their uniform up-regulation of biosynthetic machinery, the T cells in this early activation culture represent a mixed population with varying degrees of activation and/or regulatory potential.

We observe that for some genes (e.g. *Plrg1*), changes in variability are driven by a small number of outlier cells with high expression. The interpretation of these results is not trivial as it could reflect very subtle sub-structure or genuine changes in variability. To explore this, we performed the following synthetic experiment. We artificially created a mixed population of cells by combining 5 activated $CD4^+$ T cells with a population of

93 naive CD4$^+$ T cells (see **STAR Methods**). Subsequently, we performed a differential testing (mean and residual over-dispersion) between this mixed population and a *pure* population of 93 naive CD4$^+$ T cells. As expected, this analysis shows an overall increase in variability in the mixed population. For example, among the genes that exhibit higher mean expression and higher residual over-dispersion in the mixed population, we found *Il2* — which is up-regulated upon CD4$^+$ T cell activation (see **Figure S5C**). Moreover, we observe that the genes in this category are enriched for those that are only expressed in the 5 activated CD4$^+$ T cells (see **Figure S5D**). This result suggests that differential variability testing can potentially uncover markers for heterogeneous cell states or cell types which can provide important biological insights. However, changes in residual over-dispersion that are driven by outliers can also reflect unwanted contamination (e.g. mixed cell types), hence careful data filtering and clustering analysis should be performed prior to differential variability testing.

In summary, our approach allows us to extend the finding by Martinez-Jimenez et al. (2017), dissecting immune-response genes into two functional sets: (i) homogeneous up-regulation of biosynthetic machinery components and (ii) heterogeneous up-regulation of several immunoregulatory genes.

Expression dynamics during *in vivo* CD4$^+$ T cell differentiation

In contrast to the quick transcriptional switch that occurs within hours of naive T cell activation, transcriptional changes during cellular differentiation processes are more subtle and were found to be coupled with changes in variability prior to cell fate decisions (Richard et al., 2016; Mojtahedi et al., 2016). Here, we apply our method to study changes in expression variability during CD4$^+$ T cell differentiation after malaria infection using the dataset introduced by Lönnberg et al. (2017). In particular, we focus on samples collected 2, 4 and 7 days post-malaria infection, for which more than 50 cells are available.

To study global changes in over-dispersion along the differentiation time course, we first compared posterior estimates for the gene-specific parameter δ_i, focusing on the 126 genes for which mean expression does not change (see **Figure 6A** and **STAR Methods**). This analysis suggests that the expression of these genes is most tightly regulated at day 4, when cells are in a highly proliferative state. Moreover, between day 4 and day 7, the cell population becomes more heterogeneous. This is in line with the emergence of differentiated Th1 and Tfh cells that was observed by Lönnberg et al. (2017).

We next exploited the residual over-dispersion parameters to identify changes in variability

(irrespective of changes in mean expression) between consecutive time points (see **STAR Methods**). For example, separating these genes by whether their variability increases or decreases between time points revealed four different patterns (see **Figure 6B**). These include genes whose variability systematically increases (or decreases) as well as patterns where variability is highest (or lowest) at day 4.

In particular, differential variability analysis between day 2 and day 4, revealed changes in expression variability for a set of immune-related genes (see **Figure 6C**). For example, expression of *Cxcr5* which encodes the chemokine receptor that directs Tfh cells to the B cell follicles (Crotty, 2014), strongly increases in variability on day 4. This finding agrees with results from Lönnberg et al. (2017), where Tfh and Th1 differentiation was observed to be transcriptionally detectable at day 4 within a subset of activated cells. A similar behaviour was observed for *Tyk2* and *Tigit*. The latter encodes a receptor that is expressed by a subset of Tfh cells and that was found to promote Tfh function (Godefroy et al., 2015). In contrast, we observe a decrease in variability between day 2 and day 4 for *Ikzf4* (Treg-associated gene), *Ly6c1* (expressed by effector T cells) and *Tbx21* (encoding the Th1 lineage-defining transcription factor Tbet). Subsequently, we summarize the results of our differential testing between day 2 and day 4 as well as day 4 and day 7, focusing on genes that were previously detected to be Th1 or Tfh lineage associated (Lönnberg et al., 2017). We detected a continuous increase in expression of Th1-associated genes but not Tfh-associated genes (see **Figure S6A** and **STAR Methods**) with the majority of changes in variability for these genes occurring between day 2 and day 4.

We next examined immune-related genes (*Il2ra, Tbx21, Il2rb, Cxcr5, Selplg, Id2, Ifng, Icos, Ifngr1*) that were previously described as showing differences in their peak expression over the pseudo time-course of differentiation (see **Figure S6B**, **STAR Methods** and Lönnberg et al., 2017). From this list, the lineage-associated genes *Tbx21* and *Cxcr5* are up-regulated between days 2 and 4. However, these genes exhibit opposite behaviours in terms of variability: *Cxcr5* increases and *Tbx21* decreases in variability between day 2 and day 4 (see **Figure 6D**). The fact that variability of *Tbx21* (Tbet) expression was highest on day 2 suggests that Tbet is up-regulated very early in differentiation, as seen in Lönnberg et al. (2017) and similar to *in vitro* Th1 induction (Szabo et al., 2000). Moreover, this suggests that Th1 fate decisions (for at least a subset of cells) may be made even earlier than the differentiation bifurcation point identified on day 4 by the original study (Lönnberg et al., 2017).

Discussion

In recent years, the importance of modulating cell-to-cell transcriptional variation within cell populations for tissue function maintenance and development has become apparent (Bahar Halpern et al., 2015; Mojtahedi et al., 2016; Goolam et al., 2016). Here, we present a statistical approach to robustly test changes in expression variability between cell populations using scRNAseq data. Our method uses a hierarchical Bayes formulation to extend the BASiCS framework by addressing (increasingly popular) experimental protocols where spike-in sequences are not available and by incorporating an additional set of residual over-dispersion parameters ϵ_i that are not confounded by changes in mean expression. Together, these extensions ensure a broader applicability of the BASiCS software and allow statistical testing of changes in variability that are not confounded by technical noise or mean expression.

In general, stable gene-specific variability estimates ideally require a large and deeply sequenced dataset containing a homogeneous cell population (the use of unique molecular identifiers for quantifying transcript counts can also improve variability estimation, see Grün et al., 2014). However, we observe that the regression BASiCS model leads to more stable inference that requires fewer cells to accurately estimate gene-specific summaries, particularly for lowly expressed genes. Despite this, careful considerations should be taken in extreme scenarios where the number of cells is small and/or the data is highly sparse (e.g. droplet based approaches). These features of the data not only affect parameter estimation but also downstream differential testing. For sparse datasets with low numbers of cells, we recommend the use of a stringent minimum tolerance threshold and/or calibrating the test to a low expected false discovery rate (e.g. 1%) to avoid detecting spurious signals. Moreover, if possible, an internal calibration can be performed to find a reasonable minimum tolerance threshold (e.g. by randomly permuting cells between two groups to calibrate the null distribution of the differences between populations).

Our method allows characterisation of the extent and nature of variable gene expression in $CD4^+$ T cell activation and differentiation. Firstly, we observe that during acute activation of naive T cells, genes of the biosynthetic machinery are homogeneously up-regulated, while specific immune-related genes become more heterogeneously up-regulated. In particular, increased variability in expression of the apoptosis-inducing Fas ligand (Strasser et al., 2009) and the inhibitory ligand PD-L1 (Chikuma, 2016) suggests a mechanism by which newly activated cells might suppress re-activation of effector cells, thereby dynamically modulating the population response to activation. Likewise, more variable expression of Smad3, which translates inhibitory TGFβ signals into transcriptional changes (Delisle

et al., 2013), may indicate increased diversity in cellular responses to this signal. Increased variability in *Pou2f2* (Oct2) expression after activation suggests heterogeneous activities of the NF-κB and/or NFAT signalling cascades that control its expression (Mueller et al., 2013). Moreover, we detect up-regulated and more variable *Il2* expression, suggesting heterogeneous IL-2 protein expression, which is known to enable T cell population responses (Fuhrmann et al., 2016).

Finally, we studied changes in gene expression variability during $CD4^+$ T cell differentiation towards a Th1 and Tfh cell state over a 7 day time course after *in vivo* malaria infection (Lönnberg et al., 2017). Our analysis provides several insights into this differentiation system. Firstly, we observe a tighter regulation in gene expression among genes that do not change in mean expression during differentiation at day 4 at which divergence of Th1 and Tfh differentiation was previously identified (Lönnberg et al., 2017). This decrease in variability on day 4 is potentially due to induction of a strong pan-lineage proliferation program. However, we observe that not all genes follow this trend and uncover four different patterns of variability changes. Secondly, we observe that several Tfh and Th1 lineage-associated genes change in expression variability between days 2 and 4. For example, we noted a decrease in variability for one key Th1 regulator, *Tbx21* (encoding Tbet), which suggests that a subset of cells may have already committed to the Th1 lineage at day 2. Three additional Th1 lineage-associated genes also followed this trend (*Ahnak*, *Ctsd*, *Tmem154*). These data suggest that differentiation fate decisions may arise as early as day 2 in subpopulations within this system, resulting in high gene expression variability. Such an effect is in accordance with the early commitment to effector T cell fates that was previously observed during viral infection (Choi et al., 2011). As these results illustrate, diversity in differentiation state within a population of T cells can drive our differential variability results. To further disect these results, subsequent analyses such as the pseudotime inference used in Lönnberg et al. (2017) could be used to characterize a continuous differentiation process.

In sum, our model provides a robust tool for understanding the role of heterogeneity in gene expression during cell fate decisions. With the increasing use of scRNAseq to study this phenomenon, our and other related tools will become increasingly important.

Acknowledgements

NE was funded by the European Molecular Biology Laboratory (EMBL) international PhD programme. ACR was funded by the MRC Skills Development Fellowship (MR/P014178/1). SR was funded by MRC grant MC_UP_0801/1. JCM was funded by core support of Cancer Research UK and EMBL. CAV was funded by The Alan Turing Institute, EPSRC grant EP/N510129/1. We thank Dominic Grün for kindly providing access to the smFISH dataset through personal communication. Moreover, we thank Aaron Lun and Michael Morgan for constructive discussions on earlier versions of this manuscript.

Author Contributions

N.E., J.C.M., and C.A.V. designed the study; N.E. and C.A.V build the model and performed computational analyses; N.E. and A.C.R performed biological data interpretation; S.R. provided statistical assistance; and N.E., A.C.R., J.C.M., and C.A.V. wrote the manuscript. All authors commented on and approved the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

Anders, S., Pyl, P. T., and Huber, W. (2014), HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England) *31*, 166–169.

Antolović, V., Miermont, A., Corrigan, A. M., and Chubb, J. R. (2017), Generation of Single-Cell Transcript Variability by Repression. Current Biology *27*, 1811–1817.

Araki, K., Morita, M., Bederman, A. G., Konieczny, B. T., Kissick, H. T., Sonenberg, N., and Ahmed, R. (2017), Translation is actively regulated during the differentiation of CD8 + effector T cells. Nature Immunology *18*, 1046–1057.

Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A., and Itzkovitz, S. (2015), Bursty gene expression in the intact mammalian liver. Molecular Cell *58*, 147–156.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006), Noise in protein expression scales with natural protein abundance. Nature Genetics *38*, 636–643.

Bochkina, N. and Richardson, S. (2007), Tail posterior probability for inference in pairwise and multiclass gene expression data. Biometrics *63*, 1117–1125.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. a., Marioni, J. C., et al. (2013), Accounting for technical noise in single-cell RNA-seq experiments.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/24056876*

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015), Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–90.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. a., Marioni, J. C., and Stegle, O. (2015), Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology *33*, 1–32.

Carroll, R. J. (1998), *Measurement Error in Epidemiologic Studies*. John Wiley & Sons, Ltd, Chichester, UK.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008), Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544–547.

Chikuma, S. (2016), Basics of PD-1 in self-tolerance, infection, and cancer immunity. International Journal of Clinical Oncology *21*, 448–455.

Choi, Y. S., Kageyama, R., Eto, D., Escobar, T. C., Johnston, R. J., Monticelli, L., Lao, C., and Crotty, S. (2011), ICOS Receptor Instructs T Follicular Helper Cell versus Effector Cell Differentiation via Induction of the Transcriptional Repressor Bcl6. Immunity *34*, 932–946.

Crotty, S. (2014), T Follicular Helper Cell Differentiation, Function, and Roles in Disease. Immunity *41*, 529–542.

Delisle, J.-S., Giroux, M., Boucher, G., Landry, J.-R., Hardy, M.-P., Lemieux, S., Jones, R. G., Wilhelm, B. T., and Perreault, C. (2013), The TGF-β-Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. Genes and Immunity *14*, 115–126.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H., and Lempicki, R. A. (2003), DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology *4*, R60.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002), Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Fernandez, C. and Steel, M. F. J. (1999), Multivariate student-t regression models: Pitfalls and inference. Biometrika *86*, 153–167.

Fuhrmann, F., Lischke, T., Gross, F., Scheel, T., Bauer, L., Kalim, K. W., Radbruch, A., Herzel, H., Hutloff, A., and Baumgrass, R. (2016), Adequate immune response ensured by binary IL-2 and graded CD25 expression in a murine transfer model. eLife *5*, 1–17.

Godefroy, E., Zhong, H., Pham, P., Friedman, D., and Yazdanbakhsh, K. (2015), TIGIT-positive circulating follicular helper T cells display robust B-cell help functions: Potential role in sickle cell alloimmunization. Haematologica *100*, 1415–1425.

Goolam, M., Scialdone, A., Graham, S. J. L., MacAulay, I. C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J. C., and Zernicka-Goetz, M. (2016), Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell *165*, 61–74.

Grün, D., Kester, L., and van Oudenaarden, A. (2014), Validation of noise models for single-cell transcriptomics. Nature methods *11*, 637–40.

Hornung, G., Bar-ziv, R., Rosin, D., Tokuriki, N., Tawfik, D. S., Oren, M., and Barkai, N. (2012), Noise-mean relationship in mutated promoters. Genome research *22*, 2409–2417.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011), Synthetic spike-in standards for RNA-seq experiments. Genome Research *21*, 1543–1551.

Kapourani, C. A. and Sanguinetti, G. (2016), Higher order methylation features for clustering and prediction in epigenomic studies. Bioinformatics *32*, i405–i412.

Keren, L., Van Dijk, D., Weingarten-Gabbay, S., Davidi, D., Jona, G., Weinberger, A., Milo, R., and Segal, E. (2015), Noise in gene expression is coupled to growth rate. Genome Research *25*, 1893–1902.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015), Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., Tuck, A. C., Gao, X., Bühler, M., Liu, P., et al. (2015), Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell *17*, 471–485.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016), A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology *17*, 222.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U. N., et al. (2017), Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves Th1/Tfh fate bifurcation in malaria. Science Immunology *2*, eaal2192.

Love, M. I., Huber, W., and Anders, S. (2014), Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology *15*, 1–21.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016), Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology *17*, 75.

Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Göttgens, B., and Marioni, J. C. (2017), Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. Genome research *27*, 1795–1806.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015), Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Martinez-Jimenez, C. P., Eling, N., Chen, H.-c., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J. T., Teichmann, S. A., et al. (2017), Aging increases cell-to-cell transcriptional variability upon immune stimulation. Science *1436*, 1433–1436.

Miller, K. S. (1981), On the Inverse of the Sum of Matrices. Mathematics Magazine *54*, 67–72.

Mohammed, H., Hernando-Herraez, I., Savino, A., Nichols, J., Marioni, J. C., Reik, W., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., et al. (2017), Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. Cell Reports *20*, 1215–1228.

Mojtahedi, M., Skupin, A., Zhou, J., Castano, I. G., Leong-Quong, R. Y. Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016), Cell fate decision as high-dimensional critical state transition. PLoS Biology *14*, 1–28.

Mueller, K., Quandt, J., Marienfeld, R. B., Weihrich, P., Fiedler, K., Claussnitzer, M., Laumen, H., Vaeth, M., Berberich-Siebelt, F., Serfling, E., et al. (2013), Octamer-dependent transcription in T cells is mediated by NFAT and NF-iB. Nucleic Acids Research *41*, 2138–2154.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006), Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics *5*, 155–176.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015), Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell *163*, 1663–1677.

Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J., et al. (2016), Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. PLoS Biology *14*, 1–35.

Roberts, G. O. and Rosenthal, J. S. (2009), Examples of adaptive MCMC. Journal of Computational and Graphical Statistics *18*, 349–367.

Smallwood, S. a., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014), Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature methods *11*, 817–20.

Strasser, A., Jost, P. J., and Nagata, S. (2009), The many roles of FAS receptor signaling in the immune system. Immunity *30*, 180–192.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018), Exponential scaling of single-cell RNA-seq in the last decade. Nature protocols *13*, 599–604.

Szabo, S. J., Kim, S. T., Costa, G. L., Zhang, X., Fathman, C., and Glimcher, L. H. (2000), A novel transcription factor, T-bet, directs Th1 lineage commitment. Cell *100*, 655–669.

Tan, T. C. J., Knight, J., Sbarrato, T., Dudek, K., Willis, A. E., and Zamoyska, R. (2017), Suboptimal T-cell receptor signaling compromises protein translation, ribosome biogenesis, and proliferation of mouse CD8 T cells. Proceedings of the National Academy of Sciences , 201700939.

Taniguchi, Y., Choi, P. J., Li, G.-w., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. (2010), Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science (New York, N.Y.) *329*, 533–539.

Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017), Batch effects and the effective design of single-cell gene expression studies. Scientific Reports *7*.

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015), BASiCS: Bayesian analysis of single-cell sequencing data. PLOS Computational Biology *11*, e1004333.

Vallejos, C. A., Richardson, S., and Marioni, J. C. (2016), Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biology *17*.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017), Normalizing single-cell RNA sequencing data: challenges and opportunities. Nature Methods *14*, 565–571.

Vallejos, C. a. and Steel, M. F. J. (2015), Objective bayesian survival analysis using shape mixtures of log-normal distributions. Journal of the American Statistical Association *110*, 697–710.

West, M. and Harrison, J. (1989), *Bayesian forecasting and dynamic models*. Springer.

Wu, T. D. and Nacu, S. (2010), Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., Manno, G. L., Juréus, A., Marques, S., HermanyMunguba, He, L., Betsholtz, C., et al. (2015), Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

Zeng, C., Mulas, F., Sui, Y., Guan, T., Miller, N., Tan, Y., Liu, F., Jin, W., Carrano, A. C., Huising, M. O., et al. (2017), Pseudotemporal ordering of single cells reveals metabolic control of postnatal β cell proliferation. Cell Metabolism *25*, 1160–1175.e11.

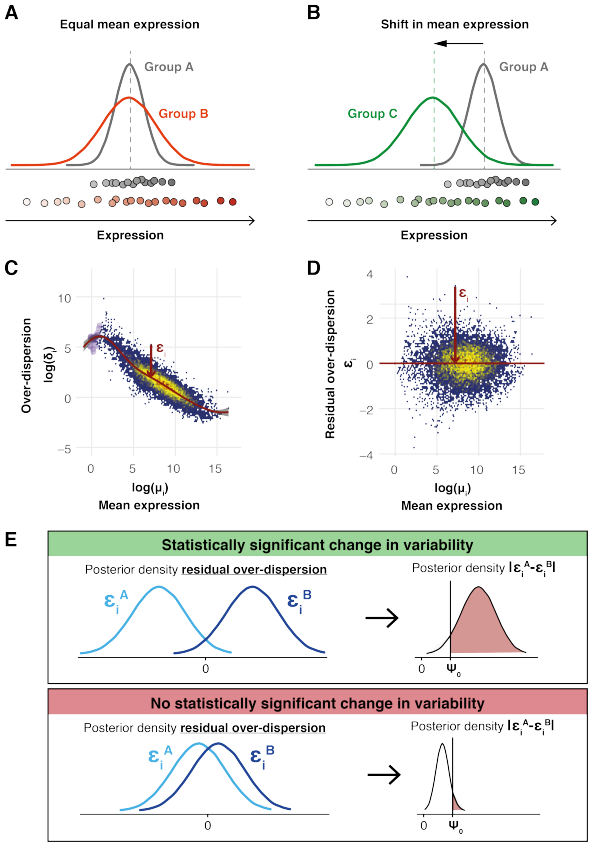

Figure 1: Avoiding the mean confounding effect when quantifying expression variability in scRNAseq data.

(A and B) Illustration of changes in expression variability for a single gene between two cell populations without (A) and with (B) changes in mean expression.

(C and D) Our extended BASiCS model infers a regression trend between gene-specific estimates of over-dispersion parameters δ_i and mean expression μ_i. Residual over-dispersion parameters ϵ_i are defined by departures from the regression trend. For a single gene, this is illustrated using a red arrow. The colour code within the scatterplots is used to represent areas with high (yellow/red) and low (blue) concentration of genes. For illustration purposes, the data introduced by Antolović et al. (2017) has been used (see **STAR Methods**).

(C) Gene-specific estimates of over-dispersion parameters δ_i were plotted against mean expression parameters μ_i. The red line shows the regression trend. This illustrates the typical confounding effect that is observed between variability and mean expression measures. Genes that are not detected in at least 2 cells are indicated by purple points.

(D) Gene-specific estimates of residual over-dispersion parameters ϵ_i were plotted against mean expression parameters μ_i. This illustrates the lack of correlation between these parameters.

(E) Illustration of how posterior uncertainty is used to highlight changes in residual over-dispersion. Two example genes with (upper panels) and without (lower panels) differential residual over-dispersion are shown. Left panels illustrate the posterior density associated to residual over-dispersion parameters ϵ_i for a gene in two groups of cells (group A: light blue, group B: dark blue). The coloured area in the right panels represents the posterior probability of observing an absolute difference $|\epsilon_i^A - \epsilon_i^B|$ that is larger than the minimum tolerance threshold ψ_0 (see **STAR Methods**).

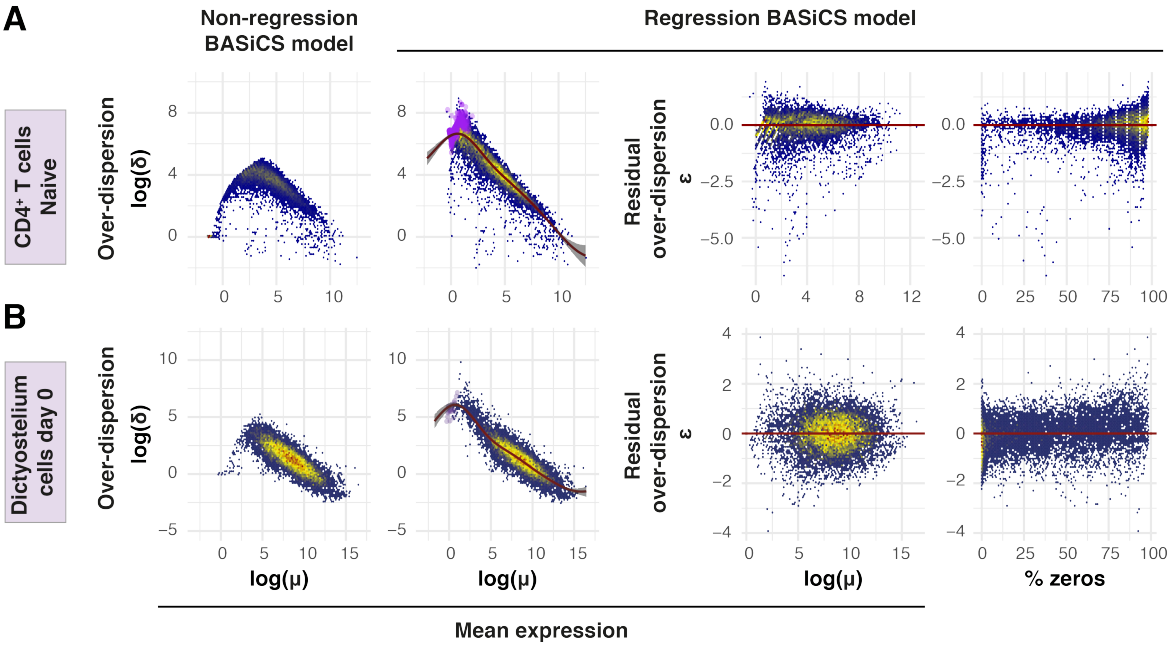

Figure 2: Parameter estimation using a variety of scRNAseq datasets.
Model parameters were estimated using the regression and non-regression BASiCS models on (A) naive CD4$^+$ T cells (Martinez-Jimenez et al., 2017) and (B) *Dictyostelium* cells prior to differentiation (day 0) (Antolović et al., 2017). These datasets were selected to highlight two situations with different levels of sparsity (i.e. the proportion of zero counts, see fourth column). More details about these datasets are provided in **STAR Methods**. The colour code within the scatterplots is used to represent areas with high (yellow/red) and low (blue) concentration of genes.
First column: gene-specific over-dispersion δ_i versus mean expression μ_i as estimated by the non-regression BASiCS model.
Second column: gene-specific over-dispersion δ_i versus mean expression μ_i as estimated by the regression BASiCS model. The red line indicates the estimated regression trend. Purple dots indicate genes detected (i.e. with at least one count) in less than 2 cells.
Third column: gene-specific residual over-dispersion ϵ_i versus mean expression μ_i as estimated by the regression BASiCS model.
Forth column: gene-specific posterior estimates for residual over-dispersion ϵ_i parameters versus percentage of zero counts for each gene. See also **Figure S2**

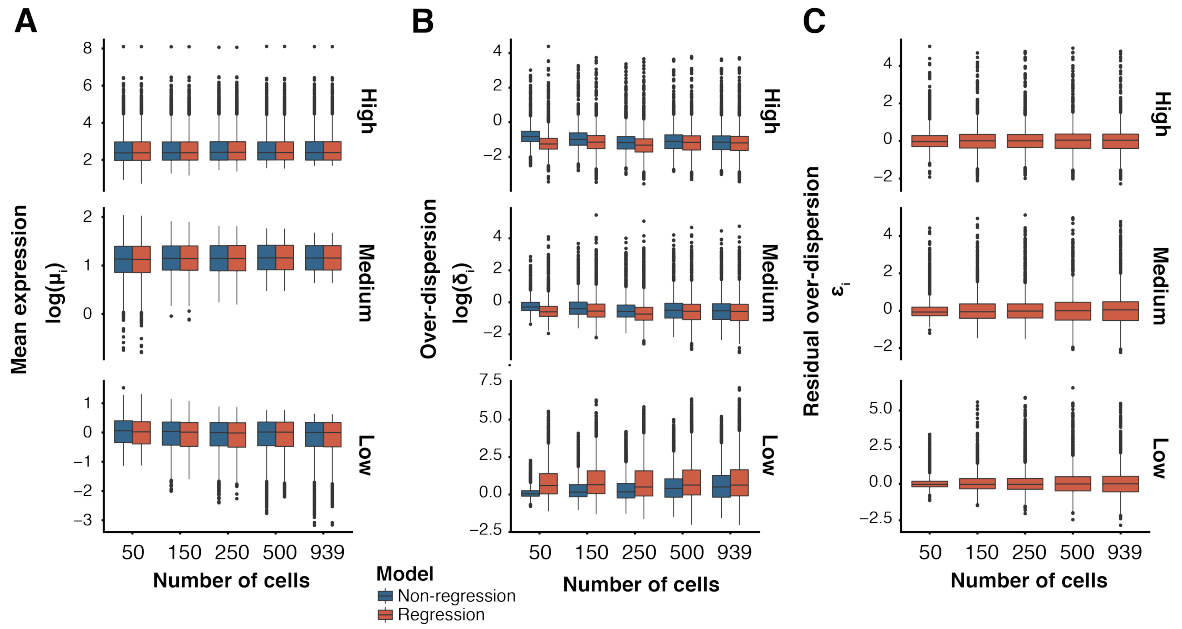

Figure 3: Estimation of gene-specific model parameters for varying sample sizes.
The regression (orange) and non-regression (blue) BASiCS models were used to estimate gene-specific model parameters for lowly (lower panels), medium (mid panels) and highly (upper panels) expressed genes across populations with varying numbers of cells. These were generated by randomly sub-sampling cells from a population of 939 CA1 pyramidal neurons (Zeisel et al., 2015). For more details see **STAR Methods**. Extended results based on multiple downsampling experiments are displayed in **Figure S3D-F**.
(A-C) For a single sub-sampling experiment, boxplots summarize the distribution of gene-specific estimates for (A) mean expression parameters μ_i (log-scale), (B) over-dispersion parameters δ_i (log-scale) and (C) residual over-dispersion parameters ϵ_i. See also **Figure S3**.

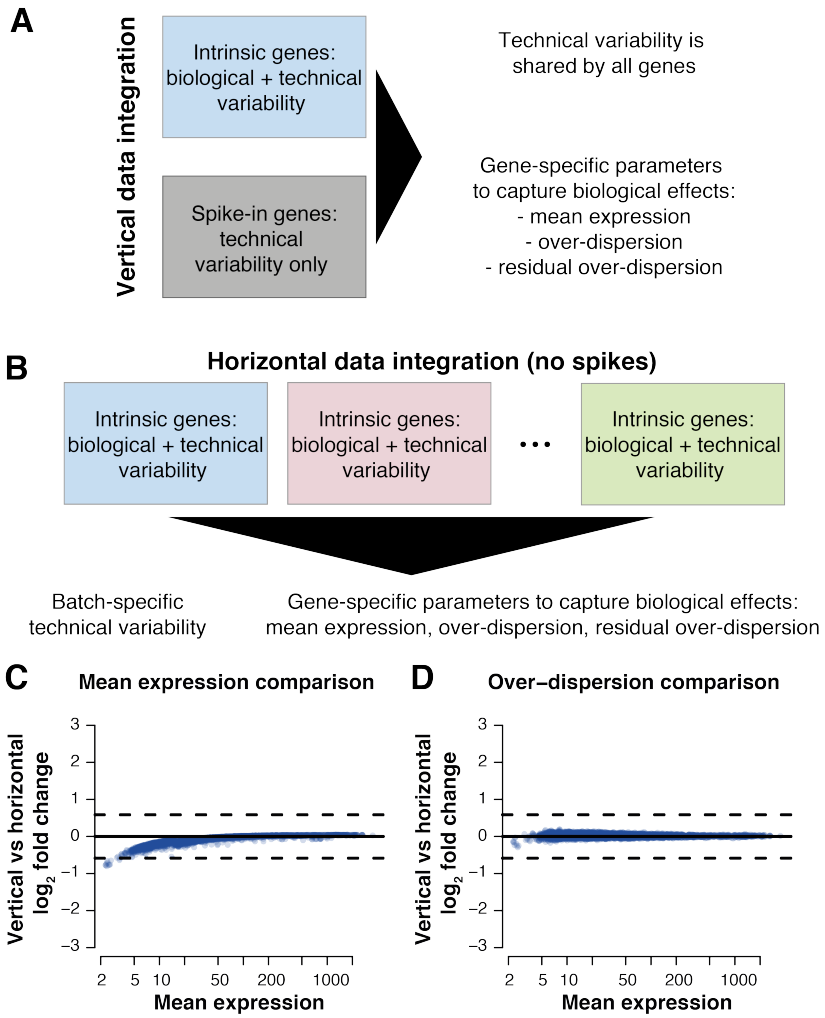

Figure 4: The spikes and no-spikes implementations of BASiCS.
(A) Diagram representing the spikes implementation of BASiCS (Vallejos et al., 2015, 2016). This uses a vertical data integration approach to borrow information from gold-standard spike-in genes to aid normalisation and to quantify technical variability.
(B) Diagram representing the no-spikes implementation of BASiCS. This uses a horizontal data integration approach to borrow information across multiple batches of sequenced cells (not confounded by the biological effect of interest) to quantify technical variability. More details about this implementation are discussed in **STAR Methods** and **Figure S4**.
(C)-(D) Comparison between the vertical and horizontal implementations of BASiCS using a dataset of mouse embryonic stem cells grown in a 2i medium (see **STAR Methods** and Grün et al., 2014). Dashed horizontal lines located at $\pm \log_2(1.5)$ indicate the default minimum tolerance $\log_2$ fold change threshold used for differential testing.
(C) Comparison in terms of posterior estimates for mean expression parameters μ_i across all genes.
(D) Comparison in terms of posterior estimates for over-dispersion parameters δ_i across all genes.
See also **Figure S4**.

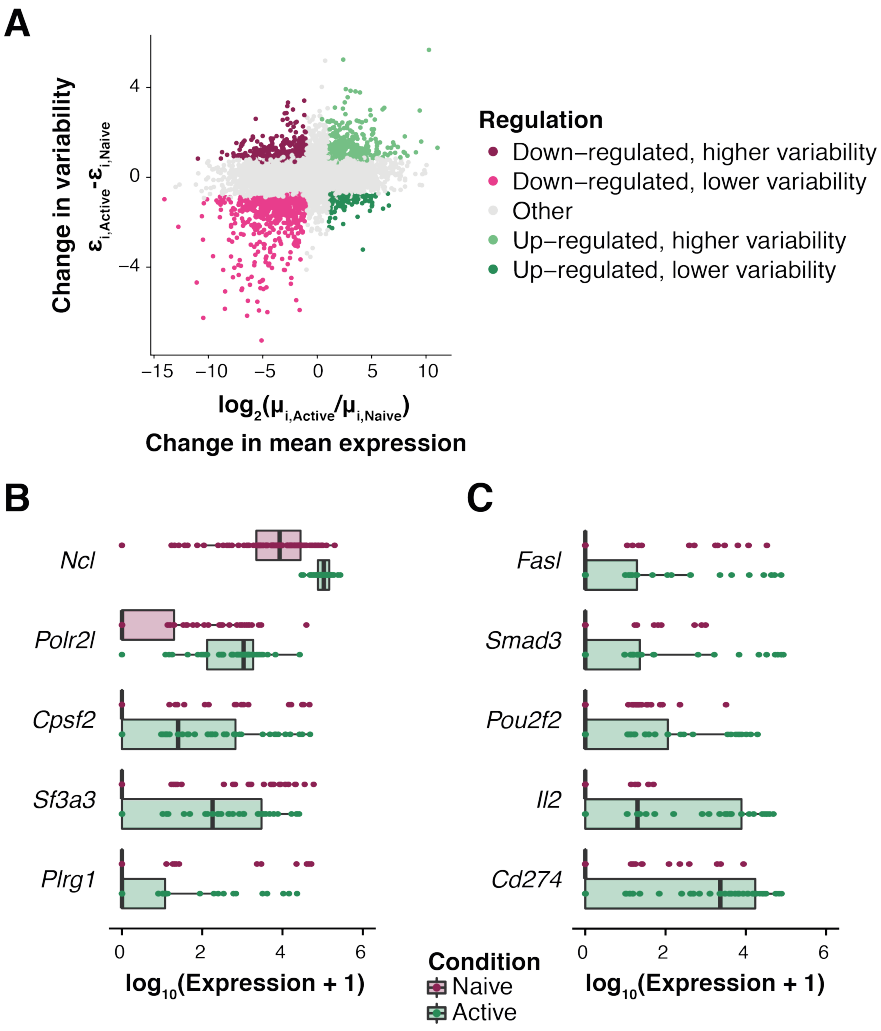

Figure 5: Changes in expression patterns during early immune activation in CD4$^+$ T cells.

Differential testing (mean and residual over-dispersion) was performed between naive and activated murine CD4$^+$ T cells. This analysis uses a minimum tolerance threshold of $\tau_0 = 1$ for changes in mean expression and a minimum tolerance threshold of $\psi_0 = 0.41$ for differential residual over-dispersion testing (expected false discovery rate is fixed at 10%, see **STAR Methods**).

(A) For each gene, the difference in residual over-dispersion estimates (Active - Naive) is plotted versus the $\log_2$ fold change in mean expression (Active/Naive). Genes with statistically significant changes in mean expression and variability are coloured based on their regulation (up/down-regulated, higher/lower variability).

(B-C) Denoised expression counts across the naive (purple) and active (green) CD4$^+$ T cell population are visualized for representative genes that (B) increase in mean expression and decrease in expression variability and (C) increase in mean expression as well as expression variability upon immune activation. Each dot represents a single cell. See also **Figure S5**.

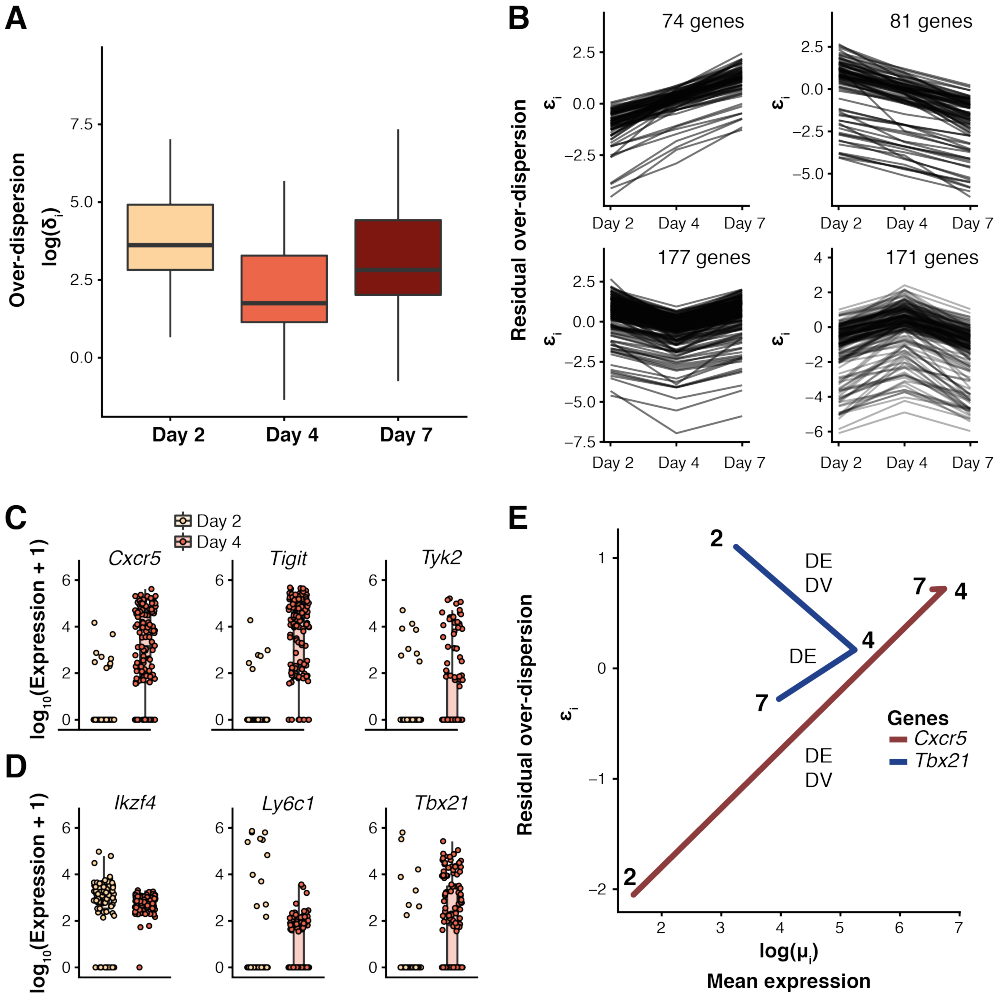

Figure 6: Dynamics of expression variability throughout CD4$^+$ T cell differentiation. Analysis was performed on CD4$^+$ T cells assayed 2 days, 4 days and 7 days after *Plasmodium* infection. Changes in residual over-dispersion were tested using a minimum tolerance threshold of $\psi_0 = 0.41$ (expected false discovery rate is fixed at 10%, see **STAR Methods**)

(A) Distribution of posterior estimates of over-dispersion parameters δ_i for genes that exhibit no changes in mean expression across the differentiation time course. Changes in mean expression were tested using a minimum tolerance threshold of $\tau_0 = 0$ (expected false discovery rate is fixed at 10%).

(B) Posterior estimates for residual over-dispersion parameters ϵ_i, focusing on genes with statistically significant changes in expression variability between time points. Gene set size is indicated for each plot.

(C-D) Denoised expression counts across cell populations at day 2 (yellow) and day 4 (red) post infection is visualized for representative genes that (C) increase or (D) decrease in variability during differentiation. Each dot represents a single cell.

(E) *Tbx21* (blue) and *Cxcr5* (red) measured at day 2, day 4 and day 7 post-infection. Posterior estimates for residual over-dispersion parameters ϵ_i are plotted against posterior estimates for mean expression parameters μ_i. Statistically significant changes in mean expression (DE, minimum tolerance threshold of $\tau_0 = 1$) and variability (DV, minimum tolerance threshold of $\psi_0 = 0.41$) are indicated for each comparison (expected false discovery rate is fixed at 10%). See also **Figure S6**.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, John Marioni (john.marioni@cruk.cam.ac.uk).

METHOD DETAILS

The BASiCS framework

The proposed statistical model builds upon BASiCS (Vallejos et al., 2015, 2016) — an integrated Bayesian framework that infers technical noise in scRNAseq datasets and simultaneously performs data normalisation as well as selected supervised downstream analyses.

Let X_{ij} be a random variable representing the expression count of gene i ($\in \{1, \ldots, q\}$) in cell j ($\in \{1, \ldots, n\}$). To control for technical noise, we employ reads from synthetic RNA spike-ins (e.g. those introduced by Jiang et al., 2011). Without loss of generality, we assume the first q_0 genes to be biological followed by the $q - q_0$ spike-in genes. As in the original BASiCS method introduced by Vallejos et al. (2015), we assume a Poisson hierarchical formulation:

$$X_{ij}|\mu_i, \phi_j, \nu_j, \rho_{ij} \overset{\text{ind}}{\sim} \begin{cases} \text{Poisson}(\phi_j \nu_j \mu_i \rho_{ij}), & i = 1, ..., q_0, j = 1, ...n; \\ \text{Poisson}(\nu_j \mu_i), & i = q_0 + 1, ..., q, j = 1, ..., n, \end{cases} \tag{1}$$

where, to account for technical and biological factors that affect the variance of the transcript counts, we incorporate two random effects:

$$\nu_j|s_j, \theta \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{1}{\theta}, \frac{1}{s_j\theta}\right) \quad \rho_{ij}|\delta_i \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{1}{\delta_i}, \frac{1}{\delta_i}\right) \tag{2}$$

In this setup, ϕ_j represents a cell-specific normalization parameter to correct for differences in mRNA content between cells and s_j models cell-specific scale differences affecting all biological and technical genes. Moreover, the random effect ν_j captures unexplained technical noise that is not accounted for by the normalisation. The strength of this noise is then quantified by a global parameter θ (shared across all genes and cells). Heterogeneous gene expression across cells is captured by ρ_{ij}, whose strength is controlled by gene-specific over-dispersion parameters δ_i. These quantify the excess of variability that

is observed with respect to Poisson sampling noise, after accounting for technical noise. Finally, gene-specific parameters μ_i represent average expression of a gene across cells.

When comparing two or more groups of cells (e.g. experimental conditions or cell types), the notation above can be extended by assuming that gene-specific parameters are also group-specific (as in Vallejos et al., 2016). Comparisons of gene-specific parameters across populations can be used to identify statistically significant changes in gene expression at the mean and the variability level. However, the well known confounding effect between mean and variability that typically arises in scRNAseq datasets (Brennecke et al., 2013) can preclude a meaningful interpretation of these results.

Modelling the confounding between mean and dispersion

Here, we extend BASiCS to account for the confounding effect described above. For this purpose, we estimate the relationship between mean and over-dispersion parameters by introducing the following joint prior distribution for $(\mu_i, \delta_i)'$:

$$\mu_i \sim \text{log-normal}\left(0, s_\mu^2\right), \qquad \delta_i|\mu_i \sim \text{log-}t_\eta\left(\text{f}(\mu_i), \sigma^2\right). \tag{3}$$

The latter is equivalent to the following non-linear regression model:

$$\log(\delta_i) = \text{f}(\mu_i) + \epsilon_i, \quad \epsilon_i \sim t_\eta(0, \sigma^2), \tag{4}$$

where $\text{f}(\mu_i)$ represents the over-dispersion (on the log-scale) that is predicted by the global trend (across all genes) expressed at a given mean expression μ_i. Therefore, ϵ_i can be interpreted as a latent gene-specific *residual over-dispersion* parameter, capturing departures from the overall trend. If a gene exhibits a positive value for ϵ_i, this indicates more variation than expected for genes with similar expression level. Accordingly, negative values of ϵ_i suggest less variation than expected for genes with similar expression level.

A similar approach was introduced by DESeq2 (Love et al., 2014) in the context of bulk RNA sequencing. Whereas DESeq2 assumes normally distributed errors when estimating this trend, here we opt for a Student-t distribution as it leads to inference that is more robust to the presence of outlier genes. Moreover, the parametric trend assumed by DESeq2 is replaced by a more flexible semi-parametric approach. This is defined by

$$f(\mu_i) = \alpha_0 + \log(\mu_i)\alpha_1 + \sum_{l=1}^{L} g_l(\log(\mu_i))\beta_l, \tag{5}$$

where $g_1(\cdot), \ldots, g_L(\cdot)$ represent a set of Gaussian radial basis function (GRBF) kernels and

$\alpha_0, \alpha_1, \beta_1, \ldots, \beta_L$ are regression coefficients. As in Kapourani and Sanguinetti (2016), these are defined as:

$$g_l(\log(\mu_i)) = \exp\left\{-\frac{1}{2}\left(\frac{\log(\mu_i) - m_l}{h_l}\right)^2\right\}, \quad l = 1, \ldots, L, \tag{6}$$

where m_l and h_l represent location and scale hyper-parameters for GRBF kernels.

In (5), the linear term captures the (typically negative) global correlation between δ_i and μ_i. Its addition also stabilises inference of GRBFs around mean expression values where only a handful of genes are observed. In (6), the location and scale hyper-parameters (m_l, h_l) are assumed to be fixed *a priori*. Details about this choice are described below.

The remaining elements of the prior were chosen as follows:

$$\begin{aligned}
\beta|\sigma^2 &\sim \mathrm{N}(m_\beta, \sigma^2 V_\beta), && (7)\\
\sigma^2 &\sim \text{Inv-Gamma}(a_{\sigma^2}, b_{\sigma^2}), && (8)\\
s_j &\overset{\text{iid}}{\sim} \text{Gamma}(a_s, b_s), \;\; j = 1, \ldots n, && (9)\\
(\phi_1, \ldots, \phi_n)' &\sim n \times \text{Dirichlet}(a_\phi), && (10)\\
\theta &\sim \text{Gamma}(a_\theta, b_\theta), && (11)
\end{aligned}$$

with all hyper-parameters fixed *a priori*. Default values are chosen as:

$$\begin{aligned}
m_\beta &= \mathbf{0}_L \text{ (an } L\text{-dimensional vector of zeroes)}, && (12)\\
V_\beta &= \mathbf{I}_L \text{ (an } L\text{-dimensional identity matrix)}, && (13)\\
a_{\sigma^2} &= 2, && (14)\\
b_{\sigma^2} &= 2, && (15)
\end{aligned}$$

with the remaining default hyper-parameter values as in Vallejos et al. (2016).

In principle, the degrees of freedom parameter η could also be estimated within a Bayesian framework. However, we observed that fixing this parameter *a priori* led to more stable results. A default choice for this parameter is described below.

Implementation

Posterior inference for the model described above is implemented by extending the Adaptive Metropolis within Gibbs sampler (Roberts and Rosenthal, 2009) that was adopted

by Vallejos et al. (2016). For this purpose, the log-Student-t distribution in 3 is represented via the same data augmentation scheme as in Vallejos and Steel (2015). The latter introduces an auxiliary set of parameters λ_i such that:

$$\delta_i|\mu_i,\beta,\sigma^2,\lambda_i,\eta \overset{\text{ind}}{\sim} \text{log-N}\left(f(\mu_i),\frac{\sigma^2}{\lambda_i}\right), \quad \lambda_i|\eta \overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{\eta}{2},\frac{\eta}{2}\right). \tag{16}$$

Moreover, the regression coefficients $\beta = (\alpha_0,\alpha_1,\beta_1,\ldots,\beta_L)'$ are inferred by noting that 5 can be rewritten as a linear regression model using

$$f(\mu_i) = X\beta, \tag{17}$$

where X is a $q_0 \times (L+2)$ matrix given by

$$X = \begin{pmatrix} 1 & \log(\mu_1) & g_1(\log(\mu_1)) & \cdots & g_L(\log(\mu_1)) \\ 1 & \log(\mu_2) & g_1(\log(\mu_2)) & \cdots & g_L(\log(\mu_2)) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \log(\mu_{q_0}) & g_1(\log(\mu_{q_0})) & \cdots & g_L(\log(\mu_{q_0})) \end{pmatrix}. \tag{18}$$

In this setting, the full conditionals associated with s_j, ϕ_j, ν_j and θ are not affected by the new prior specification of $(\mu_i,\delta_i)'$ and can be found in Vallejos et al. (2016). The full conditional for μ_i, δ_i, β, and σ^2 are derived below. As in Vallejos et al. (2015), these are derived by integrating out the random effect ρ_{ij} in 1, leading to:

$$X_{ij}|\mu_i,\delta_i,\phi_j,\nu_j \overset{\text{ind}}{\sim} \begin{cases} \text{Neg-Bin}\left(\frac{1}{\delta_i},\frac{\phi_j\nu_j\mu_i}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right), & i=1,...,q_0, j=1,...n; \\ \text{Poisson}(\nu_j\mu_i), & i=q_0+1,...,q, j=1,...,n \end{cases} \tag{19}$$

Based on 19, the likelihood function therefore takes the form

$$\begin{aligned} &\left[\prod_{i=1}^{q_0}\prod_{j=1}^{n}\frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})x_{ij}!}\left(\frac{\frac{1}{\delta_i}}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\phi_j\nu_j\mu_i}{\phi_j\nu_j\mu_i+\frac{1}{\delta_i}}\right)^{x_{ij}}\right] \\ &\times\left[\prod_{i=q_0+1}^{q}\prod_{j=1}^{n}\frac{(\nu_j\mu_i)^{x_{ij}}}{x_{ij}!}\exp\{-\nu_j\mu_i\}\right]\times\left[\prod_{j=1}^{n}\frac{(s_j\theta)^{-\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})}\nu_j^{\frac{1}{\theta}-1}\exp\left\{-\frac{\nu_j}{s_j\theta}\right\}\right]. \end{aligned} \tag{20}$$

Let $f(\mu_i)$ be as in 5. The full conditionals associated to the mean expression parameters μ_i and over-dispersion parameters δ_i are respectively given by:

$$\pi(\mu_i|\cdot) \propto \frac{\mu_i^{\sum_{j=1}^n x_{ij}}}{\prod_{j=1}^n (\phi_j \nu_j \mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}} \exp\left\{-\frac{(\log(\mu_i))^2}{2a_\mu^2} - \frac{(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2/\lambda_i}\right\}\frac{1}{\mu_i}, \tag{21}$$

$$\pi(\delta_i|\cdot) \propto \left[\prod_{j=1}^n \frac{\Gamma(x_{ij}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})}\frac{(\frac{1}{\delta_i})^{\frac{1}{\delta_i}}}{(\phi_j\nu_j\mu_i + \frac{1}{\delta_i})^{\frac{1}{\delta_i}+x_{ij}}}\right] \exp\left\{-\frac{(\log(\delta_i) - f(\mu_i))^2}{2\sigma^2/\lambda_i}\right\}\frac{1}{\delta_i}. \tag{22}$$

Moreover, the full conditionals associated to the remaining parameters λ_i, β and σ^2 are given by

$$\lambda_i|\cdot \overset{\text{ind}}{\sim} \text{Gamma}(a^*_{\lambda_i}, b^*_{\lambda_i}), \quad i = 1, \ldots, q_0, \tag{23}$$

$$\beta|\cdot \sim N(m^*_\beta, \sigma^2 V^*_\beta), \tag{24}$$

$$\sigma^2|\cdot \sim \text{Inv-Gamma}(a^*_{\sigma^2}, b^*_{\sigma^2}), \tag{25}$$

with

$$a^*_{\lambda_i} = \frac{\eta+1}{2} \tag{26}$$

$$b^*_{\lambda_i} = \frac{1}{2}\left[\frac{1}{\sigma^2}(\log(\delta_i) - f(\mu_i))^2 + \eta\right] \tag{27}$$

$$V^*_\beta = (X'\Lambda X + V_\beta^{-1})^{-1} \tag{28}$$

$$m^*_\beta = (X'\Lambda X + V_\beta^{-1})^{-1}(X'\Lambda Y + V_\beta^{-1} m_\beta) \tag{29}$$

$$a^*_{\sigma^2} = \frac{q_0 + L + 2}{2} + a_{\sigma^2} \tag{30}$$

$$\begin{aligned} b^*_{\sigma^2} &= b_{\sigma^2} + \frac{1}{2}(Y'\Lambda Y + m'_\beta V_\beta^{-1} m_\beta + \\ &\quad (\beta - m^*_\beta)'(V^*_\beta)^{-1}(\beta - m^*_\beta) - (m^*_\beta)'(V^*_\beta)^{-1} m^*_\beta), \end{aligned} \tag{31}$$

$$\equiv b_{\sigma^2} + \frac{1}{2}(Y'\Lambda Y + m'_\beta V_\beta^{-1} m_\beta + \beta'(V^*_\beta)^{-1}\beta - 2\beta'(V^*_\beta)^{-1} m^*_\beta), \tag{32}$$

where Λ is a diagonal matrix with elements $(\lambda_1, \ldots, \lambda_{q_0})$ and $Y = (\log(\delta_1), \ldots, \log(\delta_{q_0}))'$. Finally, the full conditionals associated to the global technical noise parameter (θ) and cell-specific parameters (ϕ_j, s_j and ν_j) are defined as in Vallejos et al. (2016).

Probabilistic rule associated to the differential test

We use a probabilistic approach to identify changes in gene expression between groups of cells. Let δ_i^A and δ_i^B be the over-dispersion parameters associated to gene i in groups A and B. Following 4, the $\log_2$ fold change in over-dispersion between these groups can be

decomposed as:

$$\log_2\left(\frac{\delta_i^A}{\delta_i^B}\right) = \log_2(e) \times \left[\underbrace{f^A(\mu_i^A) - f^B(\mu_i^B)}_{\text{Mean contribution}} + \underbrace{\epsilon_i^A - \epsilon_i^B}_{\text{Residual change}}\right], \tag{33}$$

where the first term captures the over-dispersion change that can be attributed to differences between μ_i^A and μ_i^B. The second term in 33 represents the change in residual over-dispersion that is not confounded by mean expression. Based on this observation, statistically significant differences in residual over-dispersion will be identified for those genes where the tail posterior probability of observing a large difference between ϵ_i^A and ϵ_i^B exceeds a certain threshold, i.e.

$$\mathrm{P}(\mid \epsilon_i^A - \epsilon_i^B \mid > \psi_0 \mid \mathrm{Data}) > \alpha_R, \tag{34}$$

where $\psi_0 > 0$ defines a minimum tolerance threshold. As a default choice, we assume $\psi_0 = \log_2(1.5)/\log_2(e) \approx 0.41$ which translates into a 50% increase in over-dispersion. In the limiting case when $\psi_0 = 0$, the probability in (34) is equal to 1 regardless of the information contained in the data. Therefore, as in Bochkina and Richardson (2007), our decision rule is based on the maximum of the posterior probabilities associated to the one-sided hypotheses $\epsilon^A - \epsilon_i^B > 0$ and $\epsilon^A - \epsilon_i^B < 0$, i.e.

$$2 \times \max\{\pi_i, 1-\pi_i\} - 1 > \alpha_R, \;\; \text{with} \;\; \pi_i = \mathrm{P}(\epsilon_i^A - \epsilon_i^B > 0 \mid \mathrm{Data}) \tag{35}$$

In both cases, the posterior probability threshold α_R is chosen to control the expected false discovery rate (EFDR) (Newton et al., 2004). The default value for EFDR is set to 10%. As a default and to support interpretability of the results, we exclude genes that are not expressed in at least 2 cells per condition from differential variability testing.
Changes in mean and over-dispersion are highlighted using the decision rule of Vallejos et al. (2016).

To evaluate the performance of our differential test we generated synthetic data under a null model (without changes in variability) and an alternative model (with changes in variability). All datasets were generated following the BASiCS model, with parameter values used set by empirical estimates based on 98 microglia cells (see below). For this purpose, we use the *BASiCS_Sim* function. To simulate data under an alternative model, 1000 genes were randomly selected and their associated δ_i's were increased or decreased by a $\log_2$ fold change of 5. Differential testing was performed either between data simulated on the same set of parameters (null model) or between data simulated from the original

parameters and the altered parameters (alternative model). We report the EFDR (Newton et al., 2004) as well as the false positive rate (FPR) for simulations under the null model and the true positive rate (TPR) for simulations under the alternative model. Synthetic data was generated with different sample sizes, with 5 repetitions for each sample size (see **Figure S1**)

Choice of hyper-parameters

As discussed above, the degrees of freedom η, the number of GRBFs L as well as the associated hyper-parameters (m_l, h_l) are set *a priori*. Here, we explain the default values implemented in the BASiCS software. These were chosen to achieve a compromise between flexibility and shrinkage strength when applied to the datasets described in **Table S1**.

Firstly, we observed that large values of L can lead to over-fitting but that small values of L can limit the flexibility to capture non-linear relations between $\log(\delta_i)$ and $\log(\mu_i)$. Thus, as a parsimonious choice, we selected $L = 10$. Moreover, as in Kapourani and Sanguinetti (2016), values for m_l were chosen to be equally spaced across the range of $\log(\mu_i)$, i.e.

$$m_l = a + (l-1)\frac{b-a}{L-1}, \quad l = 1, \ldots, L, \tag{36}$$

where $a = \min_{i \in \{1,\ldots,q_0\}}\{\log(\mu_i)\}$ and $b = \max_{i \in \{1,\ldots,q_0\}}\{\log(\mu_i)\}$. As μ_i values are unknown *a priori*, a and b are updated every 50 MCMC iterations during burn-in (fixed thereafter). Additionally, the scale hyper-parameters h_l control the width of the GRBFs and, consequently, the locality of the regression. As a default, we set these as $h_l = c \times \Delta m$, where c is a fixed proportionality constant and Δm is the distance between consecutive values of m_l. In practice, we observed that the choice of a particular value of c is not critical, as long as narrow kernels ($c < 0.5$) are avoided. As a default, $c = 1.2$ was chosen.

The degrees of freedom η controls the tails of the distribution for the residual term in 4. This influences the shrinkage towards the global trend and the robustness against outlying observations (here, these refer to genes whose mean and over-dispersion values are far from the trend). If $\eta \geq 30$, ϵ_i approximately follows a normal distribution for which posterior inference for β is known to be sensitive to outliers. Instead, small values of η introduce heavy-tails for ϵ_i, leading to more robust posterior inference. In principle, η could be estimated within a Bayesian framework. However, this is problematic as the likelihood function associated to 4 can be unbounded (Fernandez and Steel, 1999). Here, we opt for a pragmatic approach where the value of η is fixed *a priori*. To select a reasonable default value, we ran the regression BASiCS model for a grid of possible values

of η ($\eta \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 20, 25, 30\}$), using the datasets described in **Table S1** (with L, m_l and h_l fixed as described above). In all cases, we calculated a Monte Carlo estimates for the log-likelihood associated to 1 as a proxy for goodness-of-fit (data not shown). We observed that log-likelihood estimates were consistently the smallest for $\eta = 1$ and that no substantial differences are observed across larger values of η (provided that $\eta << 30$).

Based on these observations, default values implemented in the BASiCS software are set to $L = 10, c = 1.2, \eta = 5$. Despite this, the model's implementation also allows flexible adjustment of L, c and η by the user.

Running the different implementations of BASiCS

In the BASiCS R library, the default setting is to run the spikes implementation of BASiCS. The no-spikes implementation can be used by setting *WithSpikes = FALSE* in the call to *BASiCS_MCMC*. To run the regression BASiCS model, the user can set *Regression = TRUE* in the call to *BASiCS_MCMC* and *Regression = FALSE* to run the non-regression BASiCS model.

The horizontal integration approach

As seen in **Figure 4A**, BASiCS (Vallejos et al., 2015, 2016) builds upon a vertical integration framework, exploiting a set of spike-in sequences (e.g. the set of 92 ERCC molecules described in Jiang et al., 2011) as a *gold standard* to aid normalisation and to quantify technical artifacts. However, while the addition of spike-in genes prior to sequencing is theoretically appealing (Lun et al., 2017), several practical limitations affect their utility (Vallejos et al., 2017). For example, the addition of spike-ins is not trivial in droplet-based protocols such as those introduced by Klein et al. (2015) and Macosko et al. (2015).

Here, we extend BASiCS to not rely on spike-in genes using principles of measurement error models where — in the absence of gold standard features — technical variation is quantified through *replication* (Carroll, 1998). As scRNAseq is a destructive technology, it is not possible to replicate experiments by sequencing the same cells multiple times. However, we rely on the replication of population-level characteristics of the cells through appropriate experimental design (Tung et al., 2017) by randomly allocating cells from the same population to multiple independent experimental replicates (hereafter these are referred to as *batches*). Given such an experimental design, we assume that biological effects are shared across batches and that technical variation will be reflected by spurious

differences between cells and batches.

The horizontal integration model

Following this reasoning, we use a horizontal data integration approach to leverage information from multiple batches of sequenced cells to estimate biological effects that are not confounded by technical variation (see **Figure 4B**). Let X_{ijk} be a random variable representing the count (read- or UMI-based) for gene i ($\in \{1, \ldots, q\}$) in cell j ($\in \{1, \ldots, n_k\}$) of the k-th batch ($k \in \{1, \ldots, K\}$). The following model is proposed:

$$X_{ijk}|\mu_i, \nu_{jk}, \rho_{ijk} \overset{\text{ind}}{\sim} \text{Poisson}(\nu_{jk}\mu_i\rho_{ijk}), \tag{37}$$

$$\text{with } \nu_{jk}|s_{jk}, \theta \overset{\text{ind}}{\sim} \text{Gamma}(1/\theta_k, 1/(s_{jk}\theta_k)) \text{ and } \rho_{ijk}|\delta_i \overset{\text{ind}}{\sim} \text{Gamma}(1/\delta_i, 1/\delta_i). \tag{38}$$

A key assumption underlying this model is that biological effects (μ_i and δ_i) are shared across all batches and, therefore, we borrow information across cells in all batches to infer these parameters. In contrast to the original implementation of BASiCS, the absence of spike-in genes prevents the definition of two separate normalisation effects to capture nuisance differences in the scale of the observed read-counts between cells: one to capture differences in cellular mRNA content, one to capture technical artefacts (e.g. sequencing depth). Instead, in (38), the normalisation parameters s_{jk} capture a combination of these effects. The latter are inferred by borrowing information across all genes assuming that $\text{E}(s_{jk}) = 1$ *a priori*. Residual technical over-dispersion that is not captured by these normalisation parameters is captured by batch-specific parameters θ_k.

Based on the proportion of variability that is attributed to a biological component, our model can be used to identify highly and lowly variable genes within a population of cells (see Vallejos et al., 2015). Moreover, differences in mean and over-dispersion between cell populations can be highlighted by comparing gene-specific parameters (μ_i, δ_i). Finally, when adopting the prior specification described for the regression BASiCS model, our model can also be used to compare transcriptional heterogeneity in terms of a residual over-dispersion parameters ϵ_i.

Identifiability and prior specification

The model in 37-38 is not identifiable, i.e. the scale of cell-specific normalisation parameters s_{jk} and gene-specific mean expression parameters μ_i cannot be separately estimated from the data. As a solution, the following identifiability restriction is proposed:

$$\left(\prod_{i=1}^{q}\mu_i\right)^{1/q} = \mu_0 \Leftrightarrow \frac{1}{q}\sum_{i=1}^{q}\log(\mu_i) = \log(\mu_0), \quad \text{for a fixed known } \mu_0. \tag{39}$$

In 39, the geometric mean of mean expression parameters μ_i is fixed (when analysing multiple populations, this restriction independently applies within each population). In practice, we replace the value of μ_0 by its empirical counterpart, e.g. adopting the normalization strategy implemented in Lun et al. (2016). To avoid ill-defined situations, this calculation must exclude genes with zero total counts across all cells (for which the empirical estimate of μ_i is equal to 0). We note, however, that the actual value of μ_0 is not critical, as global offset effects between cell populations can be corrected post hoc (see Vallejos et al. (2016)).

Marginally, we assign a log-Normal$(0, s_\mu^2)$ prior distribution to each μ_i. However, we do not assume these parameters to be *a priori* independent. Instead, an appropriate correlation structure is introduced to satisfy the identifiability restriction in (39). Following Theorem 8.2 in West and Harrison (1989), this correlated prior is defined as

$$\log(\mu) = (\log(\mu_1), \ldots, \log(\mu_q))' \sim \mathrm{N}_q(\log(\mu_0)\mathbf{1}_q, a_\mu^2(\mathbf{I}_q - \mathbf{1}_q\mathbf{1}_q'/q)), \tag{40}$$

where q is the number of genes, $\mathbf{1}_q$ denotes a q-dimensional vector of ones and $\mathbf{I}_q$ denotes a q-dimensional identity matrix. Due to the identifiability constrain in 39, the covariance matrix in 40 is not full rank. Hence, for an arbitrarily chosen *reference* gene r, 40 can be factorised as a multivariate normal prior for
$\log(\mu_{-r}) = (\log(\mu_1), \ldots, \log(\mu_{r-1}), \ldots, \log(\mu_{r+1}), \ldots, \log(\mu_q))'$ and a point mass prior for $\log(\mu_r)|\log(\mu_{-r})$ (see Proposition 2). As a result, posterior inference can be implemented by drawing posterior samples for $\log(\mu_{-r})$, leaving posterior samples for $\log(\mu_r)$ to be completely specified by the identifiability restriction.

Using a stochastic reference gene

The vertical integration version of BASiCS (with spike-ins) is used as a benchmark for the model in 37-38. To illustrate its performance, we use the dataset of Grün et al. (2014), for which technical spike-ins and multiple batches of sequenced cells are available. In both cases, the MCMC sampler was run for 20,000 iterations, storing draws every 10 iterations and ignoring an initial burn-in period of 10,000 iterations (hence, results are shown in terms of 1,000 iterations).

Overall, posterior inference is unaffected for the majority of genes (**Figure 4C-D**). However, as it can be expected, the effect of the prior is more prominent for lowly expressed

genes where the data is less informative. In those cases, the identifiability constrain in 39 slightly shrinks posterior estimates of mean expression parameters μ_i towards μ_0. We observe that posterior inference is distorted for the arbitrarily chosen reference gene (see **Figure S4A-B**). To overcome this problem, we introduce the use of a stochastic reference choice. The latter randomly selects a reference gene at each iteration of the MCMC algorithm. As a result, each gene is treated as reference only a small proportion of times, leading to valid posterior inference for all genes (see **Figure S4C-D**).

Technical details

A correlated prior to satisfy the identifiability restriction

Proposition 1. *The prior distribution*

$$\mu_i \sim \textit{log-}N(0, a_\mu^2), \quad \textit{subject to} \quad \left(\prod_{i=1}^{q} \mu_i\right)^{1/q} = \mu_0 \quad (\textit{for fixed } \mu_0) \tag{41}$$

is equivalent to

$$\log(\mu) = (\log(\mu_1), \ldots, \log(\mu_q))' \sim N_q(\log(\mu_0)\mathbf{1}_q, a_\mu^2(\mathbf{I}_q - \mathbf{1}_q\mathbf{1}_q'/q)), \tag{42}$$

where $\mathbf{1}_q$ denotes a q-dimensional vector of ones and $\mathbf{I}_q$ denotes a q-dimensional identity matrix.

Proof. The proof follows the same steps as in the proof of Theorem 8.2 in West and Harrison (1989). Let $M = \sum_{i=1}^{q} \log(\mu_i)$. It can be shown that

$$\begin{aligned} \begin{pmatrix} \log(\mu) \\ M \end{pmatrix} \sim & \quad \mathrm{N}_{q+1}\left(\mathbf{0}_{q+1}, \begin{pmatrix} a_\mu^2\mathbf{I}_q & a_\mu^2\mathbf{1}_q \\ a_\mu^2\mathbf{1}_q' & a_\mu^2\mathbf{1}_q'\mathbf{1}_q \end{pmatrix}\right) \\ & \equiv \mathrm{N}_{q+1}\left(\mathbf{0}_{q+1}, \begin{pmatrix} a_\mu^2\mathrm{I}_q & a_\mu^2\mathbf{1}_q \\ a_\mu^2\mathbf{1}_q' & a_\mu^2 q \end{pmatrix}\right) \end{aligned} \tag{43}$$

Hence

$$\log(\mu)|M \sim \mathrm{N}_q\left((M/q)\mathbf{1}_q, a_\mu^2(\mathrm{I}_q - \mathbf{1}_q\mathbf{1}_q'/q)\right). \tag{44}$$

Finally, replacing $M = q\log(\mu_0)$, we obtain

$$\log(\mu)|\,(M = q\log(\mu_0)) \sim \mathrm{N}_q(\log(\mu_0)\mathbf{1}_q, a_\mu^2(\mathrm{I}_q - \mathbf{1}_q\mathbf{1}_q'/q)). \tag{45}$$

□

Proposition 2. *Let* $\log(\mu_{-r}) \equiv (\log(\mu_1), \ldots, \log(\mu_{r-1}), \log(\mu_{r+1}), \ldots, \log(\mu_q))'$*, where* r $(1 \leq r \leq q)$ *denotes an arbitrarily chosen reference gene. The correlated prior derived in Proposition 1 can be factorized in terms of a multivariate normal prior for* $\log(\mu_{-r})$ *and a point mass prior for* $\log(\mu_r)|\log(\mu_{-r})$ *which is located at* $q\log(\mu_0) - \sum_{i\neq r}\log(\mu_i)$*.*

Proof. Standard multivariate normal theory leads to

$$\log(\mu_{-r}) \sim \mathrm{N}_{q-1}(\log(\mu_0)\mathbf{1}_{q-1}, a_\mu^2(\mathrm{I}_{q-1} - \mathbf{1}_{q-1}\mathbf{1}'_{q-1}/q)) \tag{46}$$

and

$$\log(\mu_r)|\log(\mu_{-r}) \sim \mathrm{N}_1(m, \Sigma), \tag{47}$$

with

$$\begin{aligned}
m &= \log(\mu_0) + (-\mathbf{1}'_{q-1}/q)\left(\mathrm{I}_{q-1} - \mathbf{1}_{q-1}\mathbf{1}'_{q-1}/q\right)^{-1}(\log(\mu_{-r}) - \log(\mu_0)\mathbf{1}_{q-1}) \\
&= \log(\mu_0) - \mathbf{1}'_{q-1}\left(\mathrm{I}_{q-1} + \mathbf{1}_{q-1}\mathbf{1}'_{q-1}\right)(\log(\mu_{-r}) - \log(\mu_0)\mathbf{1}_{q-1})/q \\
&\qquad\qquad \text{(see Miller, 1981)} \\
&= \log(\mu_0) - q\mathbf{1}'_{q-1}(\log(\mu_{-r}) - \log(\mu_0)\mathbf{1}_{q-1})/q \\
&= \log(\mu_0) - \sum_{i\neq r}\log(\mu_i) + (q-1)\log(\mu_0) \\
&= q\log(\mu_0) - \sum_{i\neq r}\log(\mu_i)
\end{aligned} \tag{48}$$

and

$$\begin{aligned}
\Sigma &\propto (1 - 1/q) - (-\mathbf{1}'_{q-1}/q)\left(\mathrm{I}_{q-1} - \mathbf{1}_{q-1}\mathbf{1}'_{q-1}/q\right)^{-1}(\mathbf{1}_{q-1}/q) \\
&= \left(1 - \frac{1}{q}\right) - \frac{1}{q^2}\mathbf{1}'_{q-1}\left(\mathrm{I}_{q-1} + \mathbf{1}_{q-1}\mathbf{1}'_{q-1}\right)\mathbf{1}_{q-1} \text{ (see Miller, 1981)} \\
&= \left(1 - \frac{1}{q}\right) - \frac{1}{q^2}\mathbf{1}'_{q-1}\left(\mathbf{1}_{q-1} + (q-1)\mathbf{1}_{q-1}\right) \\
&= \left(1 - \frac{1}{q}\right) - \frac{1}{q^2}q\mathbf{1}'_{q-1}\mathbf{1}_{q-1} \equiv 0
\end{aligned} \tag{49}$$

□

Proposition 3. *Under the same assumptions as in Proposition 1. Let* $\mu_{-i,r}$ *is the vector obtained after removing elements* i *and* r *from* $(\mu_1, \ldots, \mu_q)'$*. It can be shown that*

$$\log(\mu_i)|\log(\mu_{-i,r}) \sim N\left(\frac{1}{2}\left(q\log(\mu_0) - \mathbf{1}'_{q-2}\log(\mu_{-i,r})\right), \frac{1}{2}a_\mu^2\right), \tag{50}$$

where $\mathbf{1}_{q-2}$ *denotes a* $(q-2)$*-dimensional vector of ones.*

Proof. Standard multivariate normal theory leads to

$$\log(\mu_i)|\log(\mu_{-i,r}) \sim \mathrm{N}_1(m, \Sigma), \tag{51}$$

with

$$\begin{aligned} m &= \log(\mu_0) + (-\mathbf{1}'_{q-2}/q)\left(\mathrm{I}_{q-2} - \mathbf{1}_{q-2}\mathbf{1}'_{q-2}/q\right)^{-1}(\log(\mu_{-i,r}) - \log(\mu_0)\mathbf{1}_{q-2}) \\ &= \log(\mu_0) - \mathbf{1}'_{q-2}\left(\mathrm{I}_{q-2} + \frac{1}{2}\mathbf{1}_{q-2}\mathbf{1}'_{q-2}\right)(\log(\mu_{-i,r}) - \log(\mu_0)\mathbf{1}_{q-2})/q \\ &\qquad\qquad \text{(see Miller, 1981)} \\ &= \log(\mu_0) - \frac{1}{2}(\mathbf{1}'_{q-2}\log(\mu_{-i,r}) - (q-2)\log(\mu_0)) \\ &= \frac{q}{2}\log(\mu_0) - \frac{1}{2}\mathbf{1}'_{q-2}\log(\mu_{-i,r}) \end{aligned} \tag{52}$$

and

$$\begin{aligned} \Sigma &= a_\mu^2\left((1-1/q) - (-\mathbf{1}'_{q-2}/q)\left(\mathrm{I}_{q-2} - \mathbf{1}_{q-2}\mathbf{1}'_{q-2}/q\right)^{-1}(\mathbf{1}_{q-2}/q)\right) \\ &= a_\mu^2\left(\left(1-\frac{1}{q}\right) - \frac{1}{q^2}\mathbf{1}'_{q-2}\left(\mathrm{I}_{q-2} + \frac{1}{2}\mathbf{1}_{q-2}\mathbf{1}'_{q-2}\right)\mathbf{1}_{q-2}\right) \text{ (see Miller, 1981)} \\ &= a_\mu^2\left(\left(1-\frac{1}{q}\right) - \frac{1}{q^2}\mathbf{1}'_{q-2}\left(\mathbf{1}_{q-2} + \frac{q-2}{2}\mathbf{1}_{q-2}\right)\right) \\ &= \frac{1}{2}a_\mu^2 \end{aligned} \tag{53}$$

□

Implementation

Bayesian inference is implemented using an adaptive Metropolis within Gibbs algorithm (Roberts and Rosenthal, 2009). After integrating out the random effects ρ_{ijk}, the full conditionals required for this implementation are based on the following likelihood function:

$$\begin{aligned} &\left[\prod_{i=1}^{q}\prod_{k=1}^{K}\prod_{j=1}^{n_k}\frac{\Gamma(x_{ijk}+\frac{1}{\delta_i})}{\Gamma(\frac{1}{\delta_i})x_{ijk}!}\left(\frac{\frac{1}{\delta_i}}{\nu_{jk}\mu_i+\frac{1}{\delta_i}}\right)^{\frac{1}{\delta_i}}\left(\frac{\nu_{jk}\mu_i}{\nu_{jk}\mu_i+\frac{1}{\delta_i}}\right)^{x_{ijk}}\right] \\ &\times\left[\prod_{k=1}^{K}\prod_{j=1}^{n_k}\frac{(s_{jk}\theta_k)^{-\frac{1}{\theta_k}}}{\Gamma(\frac{1}{\theta_k})}\nu_{jk}^{\frac{1}{\theta_k}-1}\exp\left\{-\frac{\nu_{jk}}{s_{jk}\theta_k}\right\}\right]. \end{aligned} \tag{54}$$

Let r denote an arbitrarily chosen reference gene. If μ_i and δ_i are assumed to be *a priori* independent (i.e. as in Vallejos et al., 2016), the associated full conditionals for μ_i ($i \neq r$)

are given by:

$$\pi(\mu_i|\mu_{-i,r},\cdot) \propto \frac{\mu_i^{\sum_{k=1}^{K}\sum_{j=1}^{n_k} x_{ijk}}}{\prod_{k=1}^{K}\prod_{j=1}^{n_k}(\nu_{jk}\mu_i + 1/\delta_i)^{x_{ijk}+1/\delta_i}} \times \pi(\mu_i|\mu_{-i,r}), \tag{55}$$

where $\pi(\mu_i|\mu_{-i,r})$ is defined as in Proposition 3 and $\mu_{-i,r}$ is the vector obtained after removing elements i and r from $(\mu_1,\ldots,\mu_q)'$. Due to the identifiability constrain, $\mu_r|\mu_{-r} \equiv \mu_0^q \left(\prod_{i\neq r}\mu_i\right)^{-1}$ with probability 1. If a gene i ($i \neq r$) is excluded from the identifiability constrain[1], 55 becomes

$$\pi(\mu_i|\mu_{-i,r},\cdot) \propto \frac{\mu_i^{\sum_{k=1}^{K}\sum_{j=1}^{n_k} x_{ijk}}}{\prod_{k=1}^{K}\prod_{j=1}^{n_k}(\nu_{jk}\mu_i + 1/\delta_i)^{x_{ijk}+1/\delta_i}} \times \exp\left\{-\frac{1}{2a_\mu^2}(\log(\mu_i))^2\right\}\frac{1}{\mu_i}. \tag{56}$$

Under this prior, the remaining full conditionals are given by:

$$\pi(\delta_i|\cdot) \propto \frac{\delta_i^{-(n/\delta_i)}}{\Gamma^n(1/\delta_i)}\left[\prod_{k=1}^{K}\prod_{j=1}^{n_k}\frac{\Gamma(x_{ijk}+1/\delta_i)}{(\nu_{jk}\mu_i+1/\delta_i)^{x_{ijk}+1/\delta_i}}\right]\exp\left\{-\frac{1}{2a_\delta^2}(\log(\delta_i))^2\right\}\frac{1}{\delta_i}, \tag{57}$$

$$\pi(s_{jk}|\cdots) \propto (s_{jk})^{a_s-(1/\theta_k)-1}\exp\left\{-\frac{\nu_{jk}}{s_{jk}\theta_k} - s_{jk}b_s\right\}, \tag{58}$$

$$\pi(\nu_{jk}|\cdot) \propto \frac{\nu_{jk}^{\sum_{i=1}^{q} x_{ijk}+1/\theta_k-1}}{\prod_{i=1}^{q}(\nu_{jk}\mu_i+1/\delta_i)^{x_{ijk}+1/\delta_i}}\, e^{-\nu_{jk}/(\theta_k s_{jk})}, \tag{59}$$

$$\pi(\theta_k|\cdot) \propto \frac{\left(\prod_{j=1}^{n_k}(\nu_{jk}/s_{jk})\right)^{1/\theta_k}}{\Gamma^{n_k}(1/\theta_k)}\theta_k^{a_\theta-(n_k/\theta_k)-1}\, e^{-(1/\theta_k)\sum_{j=1}^{n_k}(\nu_{jk}/s_{jk})-b_\theta\theta_k}, \tag{60}$$

where $n = \sum_{k=1}^{K} n_k$. Alternatively, if the joint informative prior is adopted, 55 and 57 are respectively replaced by

$$\begin{aligned}\pi(\mu_i|\mu_{-i,r},\cdot) \propto\ & \frac{\mu_i^{\sum_{k=1}^{K}\sum_{j=1}^{n_k} x_{ijk}}}{\prod_{k=1}^{K}\prod_{j=1}^{n_k}(\nu_{jk}\mu_i + 1/\delta_i)^{x_{ijk}+1/\delta_i}} \times \pi(\mu_i|\mu_{-i,r}) \times \\ & \exp\left\{-\frac{(\log(\mu_i))^2}{2a_\mu^2} - \frac{(\log(\delta_i)-f(\mu_i))^2}{2\sigma^2/\lambda_i}\right\}\frac{1}{\mu_i}\end{aligned} \tag{61}$$

$$\begin{aligned}\pi(\delta_i|\cdot) \propto\ & \frac{\delta_i^{-(n/\delta_i)}}{\Gamma^n(1/\delta_i)}\left[\prod_{k=1}^{K}\prod_{j=1}^{n_k}\frac{\Gamma(x_{ijk}+1/\delta_i)}{(\nu_{jk}\mu_i+1/\delta_i)^{x_{ijk}+1/\delta_i}}\right] \times \\ & \exp\left\{-\frac{(\log(\delta_i)-f(\mu_i))^2}{2\sigma^2/\lambda_i}\right\}\frac{1}{\delta_i},\end{aligned} \tag{62}$$

[1] By default, genes with less than 1 count per cell (on average) are excluded.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quality filtering of single cell RNA sequencing data

We employed a range of different datasets to test the proposed methodology. These datasets were selected to cover different experimental techniques (with and without unique molecular identifiers, UMI) and to encompass a variety of cell populations. Moreover, key features of each dataset can be found in **Table S1**.

Dictyostelium cells

Antolović et al. (2017) studied changes in expression variability between 0 hours (undifferentiated), 3 hours and 6 hours of *Dictyostelium* differentiation. Raw data is available by direct download (see Data S1 in Antolović et al., 2017). Across all time-points, 5 cells were removed due to low quality. Technical spike-in genes that were not detected and biological genes with an average expression (across all cells) smaller than 1 count were removed. In total, 433 cells (131 cells and 3 batches at 0h, 157 cells and 3 batches at 3h, and 145 cells and 3 batches at 6h) and 10551 genes (88 technical and 10650 biological genes) passed filtering. We used data from the 0h time point to test the functionality of our model.

Mouse brain cells

This dataset was composed of UMI scRNAseq data of cells isolated from the mouse somatosensory cortex and hippocampal CA1 region (Zeisel et al., 2015). Raw data is available from Gene Expression Omnibus under accession code GSE60361. Prior to the analysis, we removed technical genes with 0 total counts and biological genes for which the average count across all 3007 cells was below 0.1. The groups comprising microglia cells and CA1 neurons were chosen to be analysed. For these groups, 98 cells (microglia), 939 cells (CA1 pyramidal neurons) and 10744 genes (10687 biological and 57 technical genes) were left to be analysed.

Pool-and-split RNAseq data

This UMI-based dataset provides a control experiment to assess changes in biological heterogeneity in a situation where mean expression remains unchanged across conditions. Pool-and-split samples were created by pooling 1 million mESCs grown in 2i or serum medium and splitting 20pg of RNA into aliquots. These libraries are compared against single-cell samples (mESCs) (Grün et al., 2014). Raw data is available from Gene Expression Omnibus under accession code GSE54695.

As in Grün et al. (2014), some cells were removed from the analysis due to low expression of the stem cell marker Oct4. Technical genes with 0 total counts were also removed from the analysis. Additionally, lowly expressed biological genes with fewer than 0.5 counts (on average, across all samples) were excluded. This left 258 libraries (74 single mESCs grown in 2i medium, 52 single mESCs grown in serum medium, 76 pool-and-split aliquots from cells grown in 2i medium and 56 pool-and-split aliquots from cells grown in serum medium) as well as 8924 genes (50 technical spike-ins and 8874 biological genes) for the analysis. Each condition contained 2 batches.

Matched single molecule florescence *in situ* hybridization (smFISH) data from mESCs grown in 2i and serum media were obtained from Dominic Grün (Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany) through personal communications. This smFISH experiment assayed 9 genes (*Gli1*, *Klf4*, *Notch1*, *Pcna*, *Pou5f1*, *Sohlh2*, *Sox2*, *Stag3*, *Tpx2*) in more than 70 cells per condition. We excluded *Notch1* from the analysis due to strong disagreement between smFISH and scRNAseq data of cells grown in serum medium.

$CD4^+$ T cells

Non-UMI scRNAseq data of $CD4^+$ T cells were taken from Martinez-Jimenez et al. (2017). Raw data is available from ArrayExpress under accession code E-MTAB-4888. To perform a variety of tests, naive and activated $CD4^+$ T cells from young *Mus musculus* (B6) mice were selected. Biological genes with an average count $<$ 1 and non-detected technical genes were removed from the analysis. In total, 146 cells (93 naive and 53 activated $CD4^+$ T cells) and 10553 genes (10495 biological and 58 technical genes) passed filtering. Each condition contains 2 replicates.

$CD4^+$ T cell differentiation

Non-UMI scRNAseq data were generated from $CD4^+$ T cells during differentiation towards Th1 and Tfh cell fates after *Plasmodium* infection (Lönnberg et al., 2017). Raw reads were downloaded from ArrayExpress [E-MTAB-4388] and mapped against the *Mus musculus* genome (mm10) using *gsnap* (Wu and Nacu, 2010) with default settings. Read counting was performed using *HTSeq* (Anders et al., 2014) with default settings.

Quality control was performed by removing cells with fewer than 300,000 biological reads or fewer than 600,000 technical reads at day 2. At day 4 and 7, cells with fewer than 1,000,000 biological reads were excluded from downstream analysis. Additionally, we removed genes that did not show an average detection of more than 1 read at day 2, day 3, day 4 or day

7 after infection. After applying these criteria, 376 cells (Day 0: 16 cells, Day 2: 89, Day 3: 21, Day 4: 133, Day 7: 64, Day 7 non-infected: 53) and 7899 genes (7847 biological and 52 technical) remained for analysis. Note that, due to low sample sizes, we focused our analysis on data from day 2, day 4 and day 7 post-infection.

Thresholds when assessing expression changes

Statistical assessment of changes in mean expression and residual over-dispersion was performed between datasets using the regression BASiCS model. Unless otherwise indicated, the tolerance threshold was set to $\tau_0 = \log_2(1.5) = 0.58$ for differential mean expression testing, $\omega_0 = \log_2(1.5) = 0.58$ for differential over-dispersion testing and to $\psi_0 = 0.41$ for differential residual over-dispersion testing. The expected false discovery rate was controlled to 10%. This information is also displayed in figure legends.

Functional annotation analysis

We performed functional annotation analysis using DAVID version 6.8 (Dennis et al., 2003). All genes considered for differential testing were used as background. The functional annotation clustering function in DAVID was used to cluster annotation categories based on similarity and sort them according to their enrichment score.

Stabilization of posterior inference for small sample sizes

To compare parameter estimates of the regression and non-regression model across different sample sizes, we used the CA1 pyramidal neuron population from Zeisel et al. (2015). The regression BASiCS model was first run on the full population of 939 cells to generate *pseudo* ground truth parameter estimates. Subsequently, 50, 100, 150, 200, 250, 300 and 500 cells were randomly sub-sampled from the full population prior to parameter estimation. This procedure was repeated 10 times for each sample size. Based on parameter estimates using the non-regression model, we split the genes into three sets: lowly expressed ($\mu_i < 1.89$), medium expressed ($1.89 < \mu_i < 5.37$) and highly expressed ($\mu_i > 5.37$). These cut-off values were chosen such that a third of genes classifies into each category. We dissected the results of this experiment in three ways. First, we visualize boxplots showing all estimates of gene-specific parameters for a single one sub-sampling experiment (**Figure 3**). Second, we computed the $\log_2$ fold change for estimates of gene-specific over-dispersion parameters δ_i between the regression and non-regression BASiCS models (**Figure S3A-C**). Third, for each sub-sampling experiment, sample size and gene set, we computed the median $\log_2$ fold change in μ_i and δ_i and the median difference for

ϵ_i between estimates and the *pseudo* ground truth. The median and the range of these values across 10 sub-sampling experiment is used for visualization purposes (see **Figure S3D-F**).

External validation for posterior estimates of gene-specific model parameters using matched scRNAseq and smFISH data of mouse embryonic stem cells grown in 2i and serum media (see **Table S1** and Grün et al., 2014). As in Brennecke et al. (2013), residual CV^2 values for the smFISH data, we defined by the residuals obtained after fitting a gamma generalized linear model with identity link (*glmgam.fit* of the *statmod* package in R) between the CV^2 and the reciprocal log-transformed mean transcript counts.

Changes in variability during $CD4^+$ T cell activation

Firstly, we compare the results obtained by the regression BASiCS model with respect those presented in Martinez-Jimenez et al. (2017). To allow a direct comparison of the results, the same inclusion criteria as in Martinez-Jimenez et al. (2017) is adopted, i.e. we excluded genes with low mean expression ($\mu_i < 50$) in both conditions from testing. Moreover, our minimum tolerance thresholds were also adapted to match the choices in Martinez-Jimenez et al. (2017). To detect differentially expressed genes (mean) a minimum tolerance threshold $\tau_0 = 2$ was used (see **Figure S5A**). To compare the detection of differentially over-dispersed genes, we performed differential mean expression testing using a stringent minimum tolerance threshold $\tau_0 = 0$ for both models (this is to avoid the results to be confounded by changes in mean, see upper panel in **Figure S5B**). For the 463 genes that are detected as non-differentially expressed by both models for this threshold, a total of 111 genes are detected as differentially over-dispersed by either model (minimum tolerance $\log_2$ fold change threshold $\omega_0 = \log_2(1.5) = 0.58$). Out of this set, 93 genes ($\sim$83%) are detected as differentially over-dispersed by both models (see lower panel in **Figure S5B**)).

In this article, we exclude genes whose estimated mean expression parameters μ_i was below 1 from the differential testing. Furthermore, a $\log_2$ fold change threshold $\tau_0 = 1$ was adopted for mean expression testing. Unlike the more stringent threshold used by Martinez-Jimenez et al. (2017)($\tau_0 = 2$), this choice allows us to detect more subtle changes in mean expression. Moreover, the default threshold $\psi_0 = 0.41$ was used for differential variability testing. The expected false discovery rate (EFDR) was controlled to 10%.

Genes were sorted into four categories based on their changes in variability and mean expression: down-regulated upon activation with (i) lower and (ii) higher variability, and

up-regulated with (iii) lower and (iv) higher variability (see **Figure 5A**). For each of these gene sets, functional annotation analysis was performed using all tested genes as background. The functional annotation clustering tool in DAVID (Dennis et al., 2003) was used to cluster annotation categories based on similarity and sort them according to their enrichment score. Here, we list the top 3 functional annotation clusters per gene set and their corresponding enrichment score (ES):

- **Down-regulated with lower variability:** Pleckstrin homology domain (ES = 1.57), G protein signalling (ES = 1.51), glycosidase (ES = 1.49),
- **Down-regulated with higher variability:** Ankyrin repeat-containing domain (ES = 2.19), GTPase mediated signalling (ES = 1.51), steroid biosynthesis (ES = 0.89),
- **Up-regulated with lower variability:** RNA polymerase (ES = 1.6), RNA binding (ES = 1.53), splicing (ES = 1.41),
- **Up-regulated with higher variability:** Cytokine-cytokine receptor interaction (ES = 1.65), WD40 repeat (ES = 1.22), transcription (ES = 1.18).

To visualize gene expression in individual cells, we denoised the raw expression counts using the *BASiCS_DenoisedCounts* function.

Finally, we performed a synthetic experiment to illustrate how individual cells that highly express certain genes can drive the detection of changes in variability. For this purpose, we created a mixed population of cells by combining 5 activated $CD4^+$ T cells with a population of 93 naive $CD4^+$ T cells. In this mixture, response genes are lowly expressed on average and show expression outliers in a small subset of cells. *Il2* represents a gene with statistically significant higher mean expression and higher residual over-dispersion in the mixed population (see **Figure S5C**). All genes that show increased mean expression as well as increased residual over-dispersion are visualized in **Figure S5D**.

Changes in variability during $CD4^+$ T cell differentiation

To detect changes in over-dispersion and residual over-dispersion (variability) during $CD4^+$ T cell differentiation, we performed two tests between day 2 and day 4, day 4 and day 7, and day 2 and day 7. The minimum tolerance log2FC threshold to test changes in mean expression in the first test was set to $\tau_0 = 0$, while the threshold for the second test was set to $\tau_0 = 1$. The default threshold $\psi_0 = 0.41$ was used for differential variability testing. EFDR was controlled to 10%. To visualize gene expression in individual cells, we denoised

the raw expression counts using the *BASiCS_DenoisedCounts* function.

The results of the first stringent test allows us to detect genes that do not change in mean expression between any of the three time points (126 genes). For these genes, the δ_i estimates are therefore comparable across the time points, avoiding the confounded with mean expression (see **Figure 6A**). To detect genes that show different variability patterns across the time points, we first removed all genes that are expressed in fewer than 2 cells in at least one time point. For the remaining genes, the second testing strategy was used and all genes with statistically significant changes in variability between day 2 and day 4, and day 4 and day 7 were collected (see **Figure 6B**). For analysis in **Figure 6C-D** the second testing strategy was used to detect changes in variability between day 2 and day 4.

Finally, we selected gene sets listed in Lönnberg et al. (2017) to visualize their changes in mean expression and residual over-dispersion. The first set of genes is taken from Figure 3E of the original publication, which filtered genes based on their association with the bifurcation of Th1 and Tfh differentiation. The second set of genes with sequential peak expression over pseudotime is taken from Figure 5A of the original publication, which were selected based on immunological relevance from a list of dynamic genes during *in vivo* differentiation. (see **Figure S6**).

DATA AND SOFTWARE AVAILABILITY

BASiCS is freely available as part of Bioconductor 3.7 (bioconductor.org).

The results displayed in this manuscript and its supplemental material use BASiCS version 1.1.57. All R scripts for data preparation and analysis are available at github.com/MarioniLab/RegressionBASiCS2017. This link also includes instructions to download all the publicly available datasets used throughout our analyses.

Supplemental Tables

Table S2: Differential testing results between naive and activated CD4$^+$ T cells. Related to Figure 5.

For each gene highlighted to have a statistically significant change in variability: estimated difference in residual over-dispersion parameters between active and naive CD4$^+$ T cells (*DistEpsilon*), estimated $\log_2$ fold change in mean expression parameters between active and naive CD4$^+$ T cells (*Log2FCmu*) and the result of the differential test (*Regulation*).